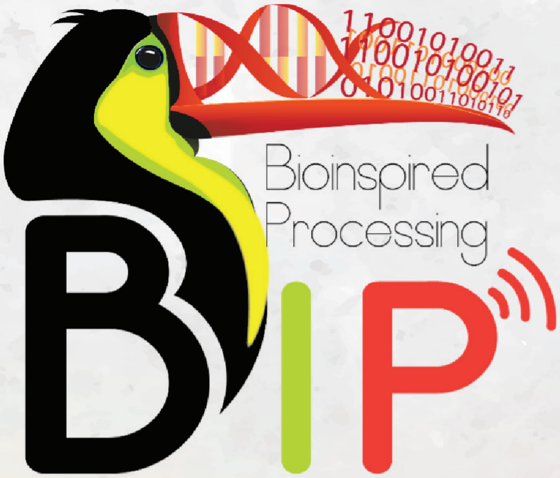
TECNOLOGÍA en marcha

Quarterly journal
August 2024
Volume 37
ISSN-E 2215-3241

Special issue
IEEE International Conference on Bioinspired Processing



Publicación y directorio en catálogos
latindex
redalyc.org UAEM
Dialnet
AmeliCA
SciELO
REDIB Red Iberoamericana de Innovación y Conocimiento Científico
DOAJ

Editorial Tecnológica de Costa Rica

TEC Tecnológico de Costa Rica

TECNOLOGÍA en marcha

Edición técnica
Alexa Ramírez Vega

Revisión filológica
Esperanza Buitrago Poveda

Diseño gráfico
Felipe Abarca Fedullo

Diagramación
Alexa Ramírez Vega

Diseño de cubierta
Ariana Sanabria García

Imagen de cubierta
Ariana Sanabria García

Datos de catalogación en publicación

Tecnología en Marcha / Editorial Tecnológica
de Costa Rica. - Vol. 37, special issue. BIP.
August, 2024 – Quarterly
ISSN-E 2215-3241

1. Ciencia y TecnologÍa –
Publicaciones periódicas CDD:600

Editorial Tecnológica de Costa Rica
TEC Tecnológico de Costa Rica

Apdo 159-7050 Cartago, Costa Rica
Tel.:(506) 2550-2297, 2550-2618
Correo electrónico: editorial@itcr.ac.cr
Web: https://www.tec.ac.cr/editorial
http://revistas.tec.ac.cr/tec_marcha

Editorial Tecnológica de Costa Rica

TEC Tecnológico de Costa Rica

La Editorial Tecnológica de Costa Rica es una dependencia especializada del Instituto Tecnológico de Costa Rica. Desde su creación, en 1978, se ha dedicado a la edición y publicación de obras en ciencia y tecnología. Las obras que se han editado abarcan distintos ámbitos respondiendo a la orientación general de la Institución.

Hasta el momento se han editado obras que abarcan distintos campos del conocimiento científico-tecnológico y han constituido aportes para los diferentes sectores de la comunidad nacional e internacional.

La principal motivación de la Editorial es recoger y difundir los conocimientos relevantes en ciencia y tecnología, llevándolos a los sectores de la comunidad que los requieren.

La revista *Tecnología en Marcha* es publicada por la Editorial Tecnológica de Costa Rica, con periodicidad trimestral. Su principal temática es la difusión de resultados de investigación en áreas de Ingeniería. El contenido de la revista está dirigido a investigadores, especialistas, docentes y estudiantes universitarios de todo el mundo.

Publicación y directorio en catálogos

redalyc.org UAEM

SciELO

latindex

amelica

Dialnet

REDIB Red Iberoamericana de Innovación y Conocimiento Científico

DOAJ DIRECTORY OF OPEN ACCESS JOURNALS

Quarterly Journal
August, 2024
Volume 37, special issue
ISSN-E: 2215-3241

TECNOLOGÍA en marcha

Contenidos

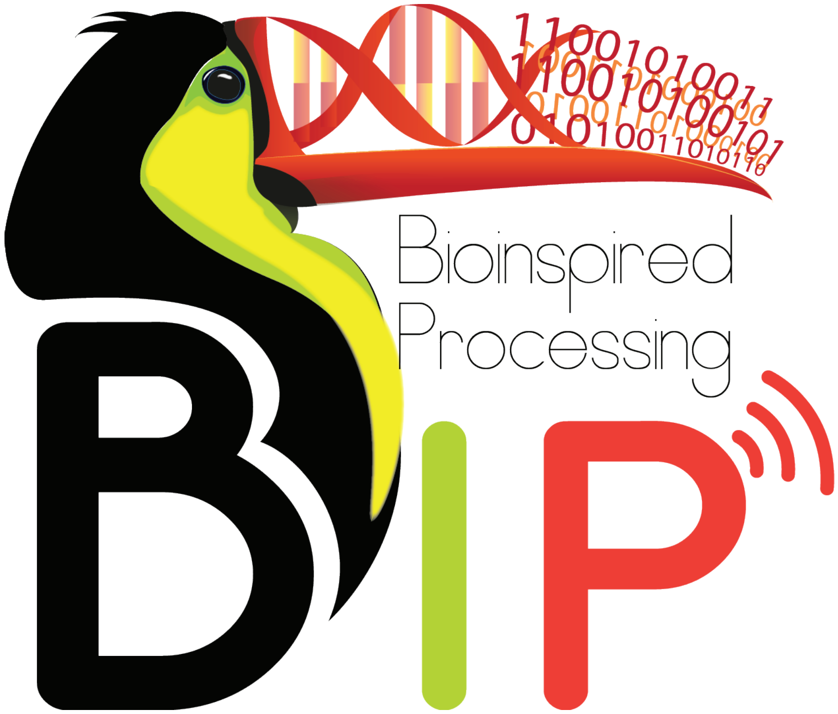
Presentation

Presentación

Andrés Segura-Castillo[1], Esteban Arias-Méndez[2],
Mauricio Rodríguez-Calvo[3], Cindy Jiménez-Picado[4]

Segura-Castillo, A; Arias-Méndez, E; Rodríguez, M; Jiménez-Picado, C. Presentation. *Tecnología en Marcha*. Vol. 37, special issue. August, 2024. IEEE International Conference on BioInspired Processing. Pag. 3-5.

https://doi.org/10.18845/tm.v37i7.7309



[1] Laboratorio de Investigación e Innovación Tecnológica (LIIT). Universidad Estatal a Distancia (UNED). Costa Rica.
[2] Profesor Investigador. Escuela de Ingeniería en Computación. Instituto Tecnológico de Costa Rica. Costa Rica.
[3] Human machine innovation Laboratory,Kanazawa University, Japón.
[4] Profesora Investigadora. Laboratorio de Investigación e Innovación Tecnológica (LIIT). Universidad Estatal a Distancia (UNED). Costa Rica.

We are delighted to present this special issue of Tecnología en Marcha, showcasing the remarkable contributions presented at the **4th and 5th IEEE International Conference on BioInspired Processing poster sessions**, held in Costa Rica in San Marcos de Tarrazú in 2022 and San Carlos in 2023, respectively. This esteemed conference has emerged as a beacon of research and innovation in the area, attracting national and international talent to share their work and insights.

Costa Rica's vibrant scientific community, coupled with its breathtaking natural landscapes, provides an ideal backdrop for fostering interdisciplinary collaborations and advancing the frontiers of bioinspired processing. As a result, the conference has become a premier platform for researchers, academics, and professionals like you to not just exchange ideas and explore emerging trends, but to shape the future of this dynamic field.

We feature a diverse range of papers that reflect the cutting-edge research and diverse perspectives presented at the conference. From bioinspired algorithms and computational models to robotics, healthcare, and beyond applications, these contributions highlight the depth and breadth of research and innovation possible in bioinspired processing.

We extend our heartfelt appreciation to each and every one of you-the authors, reviewers, organizers, and participants-who have contributed to the success of the IEEE International Conference on BioInspired Processing. It is through your individual and collective efforts that this conference continues to thrive and make significant strides in advancing scientific knowledge and technological innovation.

We hope this issue serves as a valuable resource for researchers and practitioners alike, inspiring further exploration and collaboration in the exciting field of bioinspired processing and fostering future engagement with the conference.

https://www.bipconference.org/

Nos complace presentar este número especial de Tecnología en Marcha que incluye las contribuciones presentadas en las sesiones de póster de la **4ta y 5ta Conferencia Internacional de Procesamiento Bioinspirado de IEEE**, celebradas en Costa Rica en San Marcos de Tarrazú en 2022 y San Carlos en 2023, respectivamente. Esta prestigiosa conferencia ha surgido como un faro de investigación e innovación en el área, atrayendo talento nacional e internacional para compartir su trabajo e ideas.

La activa comunidad científica de Costa Rica, junto con los paisajes naturales del país, proporcionan un telón de fondo ideal para fomentar colaboraciones interdisciplinarias y avanzar en las fronteras del procesamiento bioinspirado. La conferencia se ha convertido en un espacio de primer nivel para que personas investigadoras, académicas y profesionales intercambien ideas y exploren tendencias emergentes, contribuyen al avance futuro de este dinámico campo.

El número presenta una amplia gama de trabajos que reflejan la investigación de vanguardia y las diversas perspectivas presentadas en la conferencia. Desde algoritmos bioinspirados y modelos computacionales hasta aplicaciones en robótica y salud. Dichas contribuciones evidencian la profundidad y el potencial de la investigación e innovación posibles en el procesamiento bioinspirado.

Extendemos nuestro más sincero agradecimiento a cada una de las personas autoras, revisoras, organizadoras y participantes, que han contribuido al éxito de la Conferencia Internacional de Procesamiento Bioinspirado de IEEE. Es gracias a sus esfuerzos individuales y colectivos que esta conferencia continúa prosperando y avanzando significativamente en la promoción del conocimiento científico y la innovación tecnológica.

Esperamos que este número sirva como un recurso valioso tanto para investigadores como para profesionales, inspirando una mayor exploración y colaboración en el apasionante campo del procesamiento bioinspirado y fomentando el compromiso futuro con la conferencia.

https://www.bipconference.org/

Costa Rican fungi as potential biomaterials

Hongos costarricenses como potenciales biomateriales

Adriana Fallas-Méndez[1], Frank Solano-Campos[2], Silvia Mau-Inchaustegui[3], Giovanni Sáenz-Arce[4], Stefany Solano-González[5]

Fallas-Méndez, A; Solano-Campos, F; Mau-Inchaustegui, S; Sáenz-Arce, G; Solano-González, S. Costa rican fungi as potential biomaterials. *Tecnología en Marcha*. Vol. 37, special issue. August, 2024. IEEE International Conference on BioInspired Processing. Pag. 6-10.

https://doi.org/10.18845/tm.v37i7.7291

[1] Laboratorio de Bioinformática Aplicada, Univeridad Nacional, Costa Rica.
adriana.fallas.mendez@est.una.ac.cr
https://orcid.org/0000-0001-7812-8483

[2] Laboratorio de Biotecnología de Plantas, Universidad Nacional, Costa Rica.
frank.solano.campos@una.ac.cr
https://orcid.org/0000-0003-1055-9070

[3] Laboratorio de Biotecnología Microbiana, Universidad Nacional, Costa Rica.
silvia.mau.inchausteg@una.cr
https://orcid.org/0000-0002-9775-7442

[4] Departamento de Física, Facultad de Ciencias Exactas y Naturales, Universidad Nacional, Heredia 86-3000, Costa Rica.
Centro de Investigación en Óptica y Nanofísica, Departamento de Física, Campus Espinardo, Universidad de Murcia, 30100 Murcia, Spain.
gsaenz@una.ac.cr
https://orcid.org/0000-0003-1848-7980

[5] Laboratorio de Bioinformática Aplicada, Universidad Nacional, Costa Rica.
stefany.solano.gonzalez@una.cr
https://orcid.org/0000-0002-1167-2174

Keywords

Fungal biomaterials; mangrove; piezoelectric; DNA barcoding.

Abstract

Fungal biomaterials are gaining relevance due to their intrinsic ability of self-repair, higher sensitivity to external conditions and faster growth respective to synthetic materials. This project consists of evaluating and characterizing the physical properties of fungal strains isolated from a Pacific Coast Mangrove in Costa Rica. We identified environmental strains by recording their morphological features and complemented this by ITS-based DNA barcoding, and subsequently, classified three strains based on morphological features and seven strains by molecular analyses. Ongoing work is being done to measure electrical responses of these fungi upon light stimulation; in addition, a protocol for studying their piezoelectric properties is being developed to identify potential candidates to be used in the field of electronics. To the extent of our knowledge, our project is the first one to report piezoelectric properties from microscopic fungi in Costa Rica as means to determine its potential as biomaterials.

Palabras clave

Biomateriales fúngicos; manglar; piezoelectricidad; *barcoding*.

Resumen

Los biomateriales fúngicos han ganado relevancia en la industria debido a su capacidad intrínseca de autorreparación, mayor sensibilidad a las condiciones externas y crecimiento más rápido que los materiales sintéticos. Este proyecto consiste en evaluar y caracterizar las propiedades físicas de cepas fúngicas aisladas de un manglar de la costa del Pacífico en Costa Rica. Identificamos las cepas ambientales mediante el registro de sus características morfológicas y lo complementamos con códigos de barras de ADN basados en ITS y, posteriormente, clasificamos tres cepas en función de las características morfológicas y siete cepas mediante análisis moleculares. Se está realizando un trabajo continuo para medir las respuestas eléctricas de estos hongos ante la estimulación con luz; además, se está desarrollando un protocolo de estudio de sus propiedades piezoeléctricas para identificar posibles candidatos para ser utilizados en el campo de la electrónica. Hasta donde sabemos, nuestro proyecto es el primero en reportar propiedades piezoeléctricas de hongos microscópicos en Costa Rica como medio para determinar su potencial como biomateriales.

Introduction

Fungal organisms are considered cosmopolitans due to its wide range of colonized environments [1]. This feature is dictated by the genetic sequences within the fungal genome, to such an extent that fungi have been exploited by different industries, being recently applied as biomaterials [2] [3]. The latter covers a variety of materials useful in architecture, textile and electronics. For example, companies such as MycoWorks™ use fungal hyphae to create accessories; likewise, Ecovative Design uses fungal hyphae to develop ecological materials for construction [4] [5]. Recently, the idea of using these organisms has been expanded to the field of electronics, thanks to the intrinsic abilities of fungi, for which some authors have demonstrated its potential use in the field [6] [7]. However, to the extent of our knowledge the available reports include only mushrooms, leaving room for research to evaluate microscopic fungi potential.

Costa Rica is known for being one of the most biologically diverse countries in the world [8]; however, there are few studies directed to bioprospecting biodiversity found in different environments such as marine habitats. For this reason, the Applied Bioinformatics Laboratory (LABAP) at Universidad Nacional, focused its efforts on studying fungal strains isolated from such habitats and on utilizing their capabilities to create biotechnological products useful in a plethora of industries (e.g., biosurfactants, biomolecules and biomaterials) from both, bioinformatic and experimental approaches [9]. Herewith, we aim to study, understand and characterize Costa Rican fungal diversity and physical features as potential biomaterials by describing its piezoelectric properties.

Materials and methods

Morphological identification

In order to classify fungal isolates at the lowest possible taxonomic level, we describe fugal morphological features from potato dextrose agar (PDA) cultures. For this, we took a set of photos from every isolate in order to observe and analyze its features such as: color, geometry, texture, sporulation and presence/absence of exudates. In addition, lactophenol cotton blue was used to stain the isolates to observe hyphae septation and the type of conidiophores. This information was used to characterize morphology based on fungal taxonomic keys.

Molecular identification

In order to complement morphological identification, we carried out molecular analysis. For this, we extracted genomic DNA from every isolate using an organic solvent-based method and obtained sequences corresponding to the Internal Transcribe Spacer (ITS) region using ITS1/ITS4 primers [10]. Bidirectional Sanger sequencing was performed by Macrogen Inc. We manually reviewed each sequence using Geneious R 9.1.8 to quality control the sequences and then performed BLASTn [11] analyzes in order to identify each isolate to the genus or species level. Finally, we ran an alignment with Clustal W [12] to infer a phylogenetic tree using IQ-TREE [13], that was visualized with FirgTree [14].

Results and discussion

We have isolated 22 fungal strains from a mangrove ecosystem in the Pacific Coast of Costa Rica; from which 13 have been analyzed so far. Based on morphological traits (color, conidiophores, geometry, and shape) we found that 3 out of 13 strains are classified into *Aspergillus* and *Trichoderma* genera. The morphological classification is being complemented with molecular analysis, and so far, 7 out of 13 isolates has been taxonomically classified (Figura. 1). Identifying the isolates at the species level is vital in order to study their piezoelectric capabilities, as we aim to develop a fungal film to perform electrical measurements in response to light stimulation. To our knowledge, there are only few records which study and describe Costa Rican marine fungi [15] [16]; and therefore, this project is the first one to characterize environmental fungal strains isolated from mangrove ecosystems.

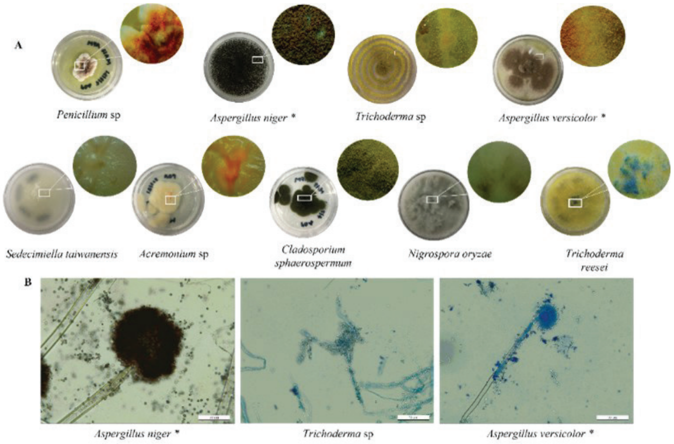


Figura. 1. Morphological record of nine marine fungal strains isolated from a mangrove environment in Costa Rica. A. Fungal strains traits on PDA. B. Fungal structures stained by lactophenol cotton blue. Species with an asterisk (*) were identified only morphologically.

The current lack of ITS sequences on public databases (such as GenBank or Uniprot) from fungal strains isolated from Costa Rican mangrove ecosystems, demonstrates the need and value to report this kind of genetic information, as this sets the required ground to get a better understanding of the organisms' biology, niche and growth requirements leading to better experimental designs. Our results display the high diversity of our isolates, clustering them in three main nodes with a good tree bootstrap (Figura. 2).

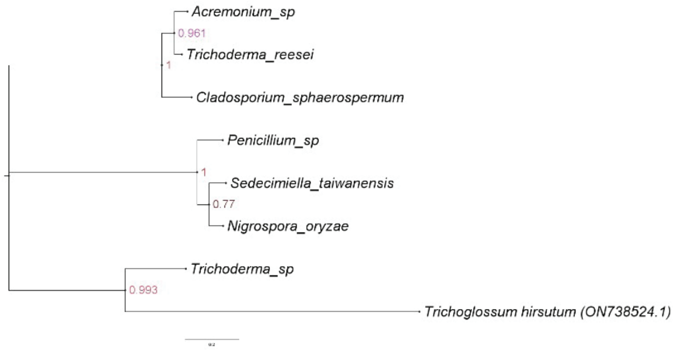


Figura. 2. Phylogenetic relationship inferred by ITS-based DNA barcoding from seven strains isolated from a mangrove ecosystem in Costa Rica. *Trichoglossum hirsutum* (ON738524.1) was used as an outgroup.

Conclusions

We have isolated, characterized, and identified fungal isolates from mangrove environmental samples which have not been reported in Costa Rica, being this the first report. In addition, we identified some methodological issues regarding the use of standard ITS1/ITS4 primers for the sequencing of the ITS region from *Aspergillus* species that consistently did not amplify. To overcome this, we propose to use CMD5/CMD6 primers, which have been tested as efficient barcoding primers for this genus.

Currently, we are testing different growth systems to obtain biofilms to further develop a protocol to evaluate piezoelectric properties of fungal strains based on atomic force microscopy analysis. Research and development of biomaterials is important due to the intrinsic characteristics of organisms that produce them, such as self-repair, higher sensitivity to external conditions and faster growth. In addition, agro-industrial residues can sustain the growth of these microorganisms, which translates into an ecological and sustainable alternative to make a variety of materials using residues as the primary source, contributing to the initiative of circular bioeconomy process. Therefore, fungal biomaterials could satisfy local industries, helping to solve current supply chain disruption related to import-export restrictions on products due to the COVID-19 pandemic.

References

[1] M. A. Naranjo-Ortiz and T. Gabaldón, "Fungal evolution: diversity, taxonomy and phylogeny of the Fungi". *Biological Reviews*, 94(6), 2101-2137, 2019.

[2] A. Gandia, J. G. van den Brandhof, F. V. Appels and M. P. Jones, "Flexible fungal materials: shaping the future". *Trends in Biotechnology*, 39(12), 1321-1331, 2021.

[3] M. Haneef, L. Ceseracciu, C. Canale, I. S. Bayer, J. A. Heredia-Guerrero and A. Athanassiou, "Advanced materials from fungal mycelium: fabrication and tuning of physical properties". *Scientific reports,* 7(1), 1-11, 2017.

[4] Ecovative Design. (2013). *Ecovative Design*. Ecovative Design. https://ecovativedesign.com/

[5] MycoWorks. (2021, July). *MycoWorks*. https://www.mycoworks.com/

[6] A. Adamatzky, A. Gandia and A. Chiolerio, "Towards fungal sensing skin". *Fungal Biology and Biotechnology*, 8(1), 2021.

[7] A. Adamatzky, A. Nikolaidou, A. Gandia, A. Chiolerio and M. M. Dehshibi, "Reactive fungal wearable". *BioSystems*, 199, 104304, 2021.

[8] J. José Alvarado, B. Herrera, L. Corrales, J. Asch and P. Paaby, "Identificación de las prioridades de conservación de la biodiversidad marina y costera en Costa Rica". *Rev. Biol. Trop.*, 2010.

[9] S. Solano-González and F. Solano-Campos, "Production of mannosylerythritol lipids: biosynthesis, multi-omics approaches and commercial exploitation". *Molecular Omics*, 2022.

[10] T. J. White, T. Bruns, S. J. W. T. Lee and J. Taylor, "Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics". Pp. 315-322 In: *PCR Protocols: A Guide to Methods and Applications*, eds. Innis, M.A., D.H. Gelfand, J.J. Sninsky, and T.J. White. Academic Press, Inc., New York, 1990.

[11] S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman, "Basic local alignment search tool". *Journal of molecular biology*, 215(3), 403-410, 1990.

[12] J. D. Thompson, D. G. Higgins and T. J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice". *Nucleic acids research*, 22(22), 4673-4680, 1994.

[13] L. T. Nguyen, H. A. Schmidt, A. Von Haeseler and B. Q. Minh, "IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies". *Molecular biology and evolution*, 32(1), 268-274, 2015.

[14] A. A. Rambaut, (2009). FigTree. Tree Figure Drawing Tool.

[15] A. Ulken, R. Víquez, C. Valiente and M. Campos, "Marine fungi (Chytridiornycetes and Thraustochytriales) frorn a rnangrove area at Punta Morales, Golfo de Nicoya, Costa Rica". *Rev. Biol. Trop.*, vol. 38, no. 2, pp. 243–250, 1990.

[16] S. Masís-Ramos, P. Meléndez-Navarro and E. Méndez-Rodríguez, "Potencial biotecnológico de los hongos marinos en las zonas costeras de Costa Rica". *Revista Tecnología en Marcha*, 34(2), 48-59, 2021.

Preliminary analysis of socioeconomic variable correlation with geospatial modeling in Costa Rica dengue epidemics

Análisis preliminar de la correlación de variables socioeconómicas con modelado geoespacial en la epidemia de dengue en Costa Rica

Cristina Soto-Rojas[1], Cesar Garita[2], Mariela Abdalah[3], Juan Gabriel Calvo[4], Fabio Sanchez[5], Esteban Meneses[6]

Soto-Rojas, C; Garita, C; Abdalah, M; Calvo, J. G; Sanchez, F; Meneses, E. Preliminary analysis of socioeconomic variable correlation with geospatial modeling in Costa Rica dengue epidemics. *Tecnología en Marcha*. Vol. 37, special issue. August, 2024. IEEE International Conference on BioInspired Processing. Pag. 11-21.

https://doi.org/10.18845/tm.v37i7.7292

[1] National High Technology Center and Costa Rica Institute of Technology. Costa Rica
csoto@cenat.ac.cr
https://orcid.org/0000-0001-9180-1628

[2] Costa Rica Institute of Technology. Costa Rica
cesar@itcr.ac.cr
https://orcid.org/0000-0003-4592-3266

[3] National High Technology Center and Costa Rica Institute of Technology. Costa Rica
mabdalah@cenat.ac.cr
https://orcid.org/0000-0002-9790-2689

[4] University of Costa Rica. Costa Rica
juan.calvo@ucr.ac.cr
https://orcid.org/0000-0001-9948-9966

[5] University of Costa Rica. Costa Rica
fabio.sanchez@ucr.ac.cr
https://orcid.org/0000-0002-5552-3672

[6] National High Technology Center and Costa Rica Institute of Technology. Costa Rica
emeneses@cenat.ac.cr
https://orcid.org/0000-0002-4307-6000

Keywords

Geospatial modeling; data science; dengue epidemics.

Abstract

Dengue is a mosquito-transmitted disease that affects more than 5 million people worldwide. It is endemic in more than 100 countries and it has presence in 5 continents. Understanding the dynamics of dengue epidemics is crucial in reducing the massive public health impact this disease has. However, dengue is a complex phenomenon. There are many variables that contribute to the spread of the virus and the interconnection of those variables is not clear. We set out to explore the correlation of socioeconomic variables in dengue epidemics by using a geospatial model. Our study is centered in Costa Rica, a country with a repeated affectation by the virus. We found a possible relationship between number of dengue cases and some socioeconomic variables (dwellings with water pipes, location of work), which open the gates to consider including them in a more sophisticated epidemiological model.

Palabras clave

Modelado geoespacial; ciencia de datos; epidemia de dengue.

Resumen

El dengue es una enfermedad transmitida por mosquitos que afecta a más de 5 millones de personas en todo el mundo. Es endémica en más de 100 países y tiene presencia en los 5 continentes. Comprender la dinámica de las epidemias de dengue es crucial para reducir el impacto masivo que tiene esta enfermedad en la salud pública. Sin embargo, el dengue es un fenómeno complejo. Hay muchas variables que contribuyen a la propagación del virus y la interconexión de esas variables no está clara. Nos propusimos explorar la correlación de las variables socioeconómicas en las epidemias de dengue mediante el uso de un modelo geoespacial. Nuestro estudio se centra en Costa Rica, país con afectación reiterada por el virus. Encontramos una posible relación entre el número de casos de dengue y algunas variables socioeconómicas (viviendas con tubería de agua, ubicación del trabajo), que abren las puertas a considerar incluirlas en un modelo epidemiológico más sofisticado.

Introduction

According to the World Health Organization, the dengue disease is endemic in more than 100 countries [1]. The affected regions are Africa, the Americas, Eastern Mediterranean, South-East Asia, and Western Pacific; being the last two the ones hit the hardest. Dengue is a disease spread by the mosquito *Aedes Aegypti* and *Aedes Albopictus*. The transmission of the virus depends on two main elements: mosquitoes and humans. The dengue mosquitoes spread a virus that belongs to the *Flaviviridae* family, and four distinct serotypes can cause the disease with different variations of symptoms. That means there are four versions of the virus that causes dengue: DENV-1, DENV-2, DENV-3, and DENV-4. If a patient recovers from one of these serotypes, the patient may gain immunity to that specific serotype, but not to the others.

The dynamics of a dengue epidemic are complex. The available literature provides several approaches to model this disease. Most of those studies use climate variables, since they can influence the mosquito habitat conditions. As human activities also impact the spread of the virus, geospatial and socioeconomic variables have also been used in those models. To understand how dengue affects Costa Rica and its different regions, we embarked on a research project to explore available datasets from multiple sources and prevalent mathematical models. The final aim of the project is to shed light on variable correlation and model robustness for dengue epidemics in the country.

Background

Geographically weighted regression model

The simple linear regression is a classic model used to describe a variable of interest, known as the dependent variable, as a linear function of independent variables known as the predictor variables. This model has been used in many scenarios, delivering accurate results. However, if it considers the independence in the predictor variables, its effectiveness still has to be evaluated for geographical analysis.

When we consider the relation between the dependent variable and the predictor variables in multiple regions, it may happen that the relationship of variables changes from region to region. Moreover, in some regions the variable could have a different impact. If we consider a simple linear regression, differences from region to region could not be captured in the model that attempts to be fitted for all, and some valuable geographical analysis could be lost.

Geographically weighted regression (GWR) [16] considers each region and their neighborhoods as an individual local regression, and as an estimation of the coefficients. Then, the results of a regression for each region and the impact of the predictor variables for each case can be observed, providing a more segregated analysis that can bring more information than a simple linear regression. The formula for an GWR is:

$$y_i = a_{i0} + \sum_{k=1}^{N} a_{ik} x_{ik} + \epsilon_i$$

where y_i is the predicted variable for the region *i*, $\mathbf{a}_{i0}$ is the local intercept, $\mathbf{a}_{ik}$ is the coefficient of the variable **k** at the region **i**, x_{ik} is the value of the independent variable for the **k** variable at the region **i**, with **N** the number of regions, and ϵ_i the error term. One of these equations is calculated for each region and the coefficient calibration considers the neighborhoods of the region.

To determine the neighborhoods of each region there are two types of strategies that were used, the fixed one and the adaptive. The fixed one considers a fixed radius and includes the points inside this circle around the central region. In extreme cases, when the regions are small or large, the fixed approach may have problems with the estimations. The adaptive one considers a proportion of observations from the nearest neighbors and includes them; it adapts providing larger bandwidths to sparse data and smaller ones on the other case.

Related work

Hwa-Lung et al. [6] created a model that allows the generation of alerts within a week considering the spatio-temporal predictions of dengue fever cases. They used a stochastic Bayesian Maximum Entropy analysis and provide valuable spatial information of the dengue fever outbreaks. Åström et al. [7] proposes a model that considers the gross domestic product per capita (GPD), obtaining that both climatic variables and GDP influence the incidence of dengue. They predict the population at risk of dengue for 2050 under different combinations of the variables. One of their conclusions is that worsening global economic conditions would contribute to increase dengue incidence, especially on vulnerable urban populations.

Lowe et al. [8] showed that including climate information in a model for the case of Thailand improved the model for 79% of the provinces they modeled. They used a Bayesian framework considering spatial and temporal variables. They modeled the dengue relative risk for the next month of the data, and found that the climate variables of temperature and precipitation have a statistically significant contribution to the relative risk in the following month. Under this relationship, many models have aimed at predicting the behavior of the dengue disease considering these climatic variables. The variables of maximum temperature, humidity, and El Niño Southern Oscillation have been mainly considered, obtaining results that predict between 1 and 4 months of this phenomenon [9]-[11].

In the case of Costa Rica, we have an optimal climate for the mosquito. Vasquez et al. [12] proposed a predictive model using a generalized additive model and random trees that allow predicting the relative risk in 5 cantons of the country. Also, Sánchez and Calvo [13] proposed an epidemic model that allows exploring the transmission dynamics of this disease in the early life stage of mosquitoes and with an age structure in humans. This model allows a better analysis of the implications of this phenomenon by providing the age distribution of humans, and collaborates with prevention to learn more about the early life stages of the mosquito.

However, although multiple models have used climatic variables, Morin et al. [14] highlighted the complexity of the relationship between climatic variables and the factors that affect dengue transmission. That situation could explain some inconsistencies in associations between climatic variables and dengue, since ecological aspects, the development of the virus and host-species interactions are usually ignored.

Delmelle et al. [15] created a GWR model that considers socioeconomic and environmental variables and found that the main influencing variables are population density, socioeconomic status, proximity to tire shops and plant nurseries, and the presence of a sewage system. Naqvi et al. [17] used a GWR model and show that the temperature is the most significantly associated variable with the dengue fever in Pakistan.

Methods

Data Sources

Dengue Cases

Incidence dengue cases were obtained from the Ministry of Health of Costa Rica, and are found at the cantonal and regional level with the following characteristics:

- Data period: 2012-2013, 2015-2019.
- Data spatial segregation: socioeconomic region (6) and canton (84).
- Available variables: Number of cases for each year by epidemiological week.
- Some regions have 0 cases in some of the years. The data is complete.

Socio-economic

Information was extracted from the National Institute of Statistics and Censuses of Costa Rica, from which the following data was extracted from the National Census:

- Data period: 2011.
- Data spatial segregation: socioeconomic region and canton.
- Available variables: telephony, access to services, household distribution, education, housing characteristics, multidimensional poverty, Gini coefficient, household income, garbage disposal system, water system access, work and mobility.
- These files contain the value for each variable per house and per person. Some present the percentage, and some the total. The data is complete.

Spatial

Various geospatial layers were obtained from the National Territorial Information System of Costa Rica. They have the following characteristics:

- Data format: geospatial object from Web Feature Service.Data spatial type: polygons.
- Available layers: cantonal distribution.
- The data is complete.

Data Wrangling

Considering the irregular behavior of the incidence of dengue, we proceeded to group all the cases to have a greater representation in each canton of the incidence. On the other hand, the socioeconomic variables used are those of the 2011 census. These data were the last available information about the state of the cantons, since the centroids of the regions on the country are closer on the central area of the country and more distant on the external area, then the data were divided on being part of the Great Metropolitan Area (GMA) or not. The GMA delimits the central regions of the country and contains the closer centroids.

Model

For the model, the variable of accumulated cases by canton was used as a dependent variable. Multiple socioeconomic variables were taken into consideration as the independent variable, but those that presented a higher correlation with the data and better results in the model are shown. Results were obtained with the variables: access to water pipes, percentage of people who work in the same canton, percentage of people who work in another canton and percentage of people who are outside the labor force. Variables that were considered but did not show significant results were: deprived population, Gini index, population working in the primary, secondary and tertiary sectors, percentage of households in slum conditions.

Considering the GWR description, y_i is the dengue cases, and x_i is the different variables consider, finally the regions consider are the cantons. A GWR model was created for each variable. To evaluate the results of the model and determine if there was a relevant contribution of each variable, the results of R2 for each model were compared. The statistical significance was evaluated for each model:

$$\mathrm{t} = \mathrm{coefficient/coefficient\_se}$$

A model was created for each variable for each case: GMA and non-GMA.

Experimental Setup

The computer used has an operating system Windows version. The code was implemented on R, version 4.2.0. The plotting library is tmap. The code used to run the model an generate the visualizations is available at the repository: https://gitlab.com/CNCA CeNAT/gwr-dengue.

Table 1. Software configuration.

Program	Version
Operating System (OS)	Windows 10 PRO
Processor	AMD Ryzen 7 2700 Eight- Core Processor 3.20 GHz
RAM	16gb

Results

Working in the same canton

This variable represents the percentage of the population in each canton that works in the same canton. The local R2 shows that the model has a modest adjustment, with values between 0.1 and 0.5 in the GMA case, and the coefficients associated show a positive correlation, especially in the upper left region (see Figure 1). In Figure 2 we observe that the model in the case of cantons outside the GMA does not show a good fit. Then it is possible to observe that the GMA seems to positively correlate people that work in the same canton with the cases.

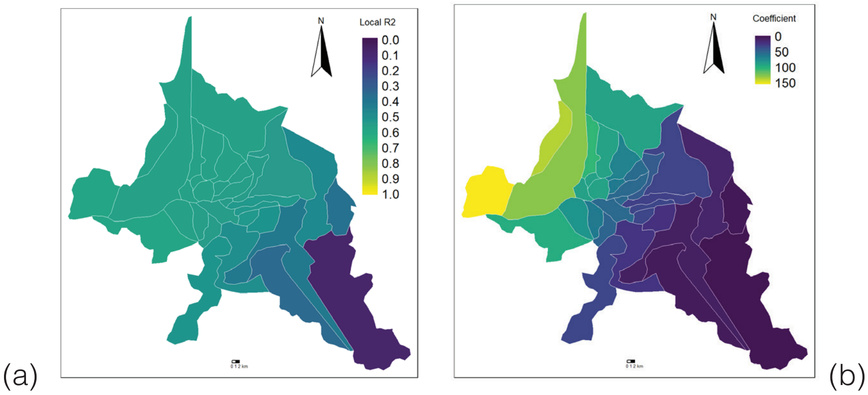


Figure 1. Results for population working on the same canton on the GMA (a) Local R2 (b) Coefficients.

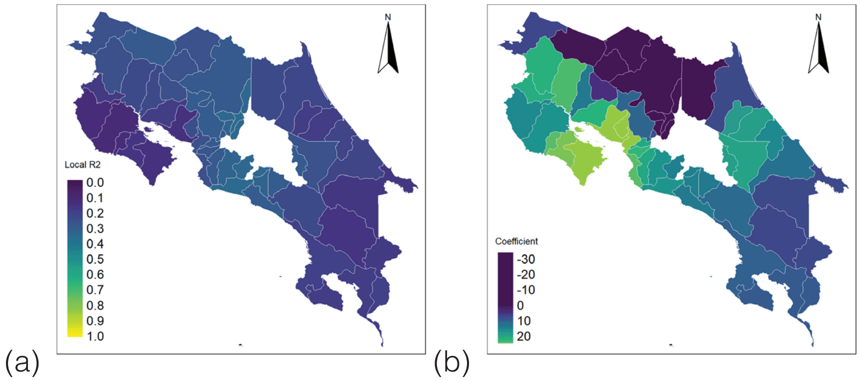


Figure 2. Results for population working on the same canton outside the GMA (a) Local R2 (b) Coefficients.

Working in another canton

This variable represents the percentage of the population that works in a different canton. The local R2 shows that the model has a modest adjustment, with values between 0.1 and 0.5 in the GMA case, and the coefficients associated show a negative correlation, especially in the upper left region (see Figure 3). It is important to note that in this case, the values of the coefficients are negative, so the model would be telling us that there may be a possible relationship between people leaving their canton to work and fewer cases of dengue.

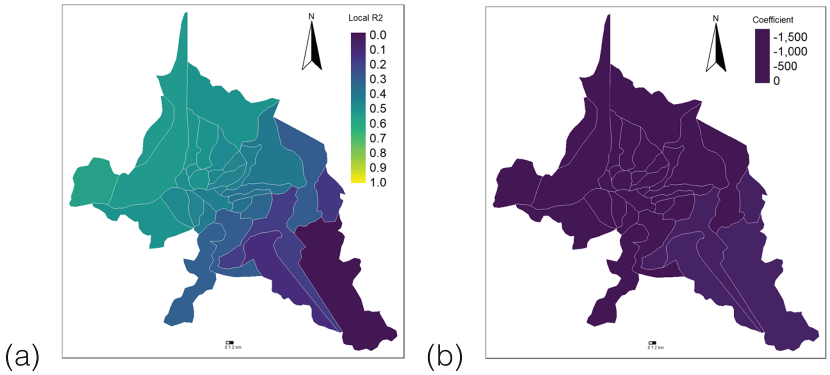


Figure 3. Results for population working in another canton on the GMA (a) Local R2 (b) Coefficients.

In the case of cantons outside the GMA, on Figure 4 we can see that, as in the case of the previous variable, robust results were not obtained.

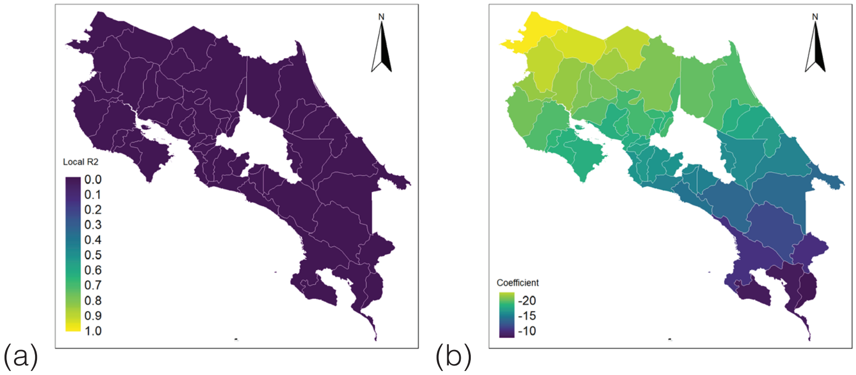


Figure 4. Results for population working in another canton outside the GMA (a) Local R2 (b) Coefficients.

Unemployed population

This variable represents the percentage of the population that is unemployed. In this case, we can see that the results follow a behavior similar to that of the two previous cases. The local R2 shows that the model has a modest adjustment, with values between 0.2 and 0.5 in the GMA case, and the coefficients associated show a positive correlation, especially in the upper left region (see Figure 5). In the GMA area, it allows us to analyze that, as with the variable of people who work in the same canton, we have positives coefficients. We could then observe a possible relationship between the incidence of dengue cases and the population that stays mainly in the same region, such as people outside the labor force together with those who work in the same canton.

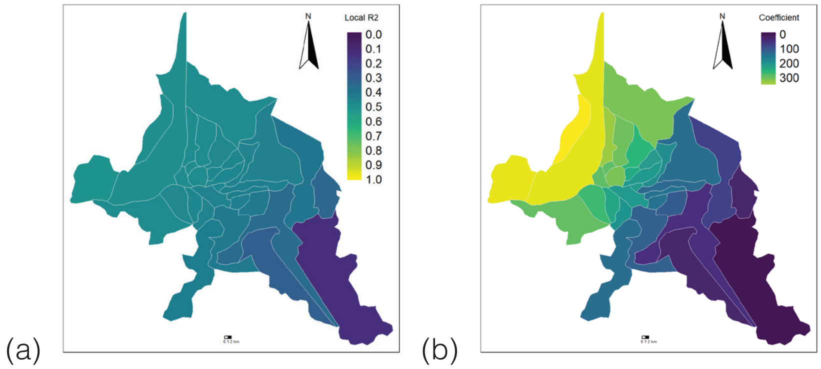


Figure 5. Results for unemployed population on the GMA (a) Local R2 (b) Coefficients.

In this case we can also observe in the Figure 6 that outside the GMA a good fit of the model was not obtained, similar to that with the other two variables.

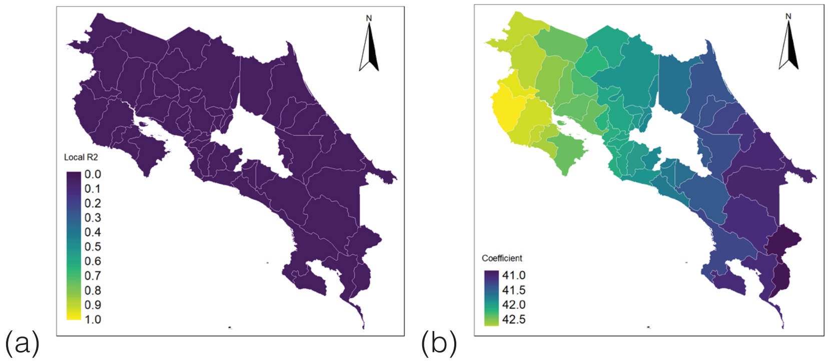


Figure 6. Results for unemployed population outside the GMA (a) Local R2 (b) Coefficients.

Water pipes

This variable represents the dwellings that do not have access to pipes. The local R2 shows that the model has a strong adjustment, with values between 0.1 and 0.9 in the GMA case, and the coefficients associated show a positive correlation, especially in the upper left region (see Figure 7). For the case of the area outside the GMA, Figure 8 shows a good fit for the model. The local R2 shows that the model has a strong adjustment, with values between 0.4 and 0.9, with coefficients that are also positive. This can be interpreted as a possible relationship between access to pipes in the house and the incidence of dengue cases.

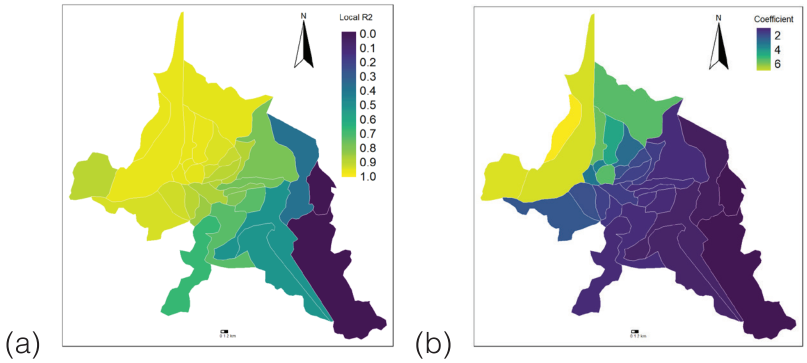


Figure 7. Results for population without water pipes at households on the GMA (a) Local R2 (b) Coefficients.

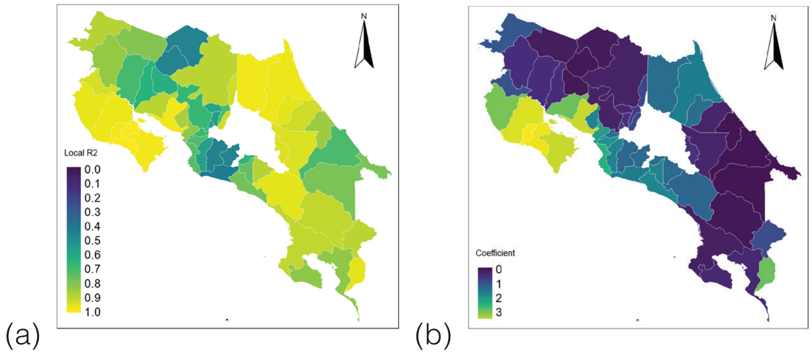


Figure 8. Results for population without water pipes at household outside the GMA (a) Local R2 (b) Coefficients.

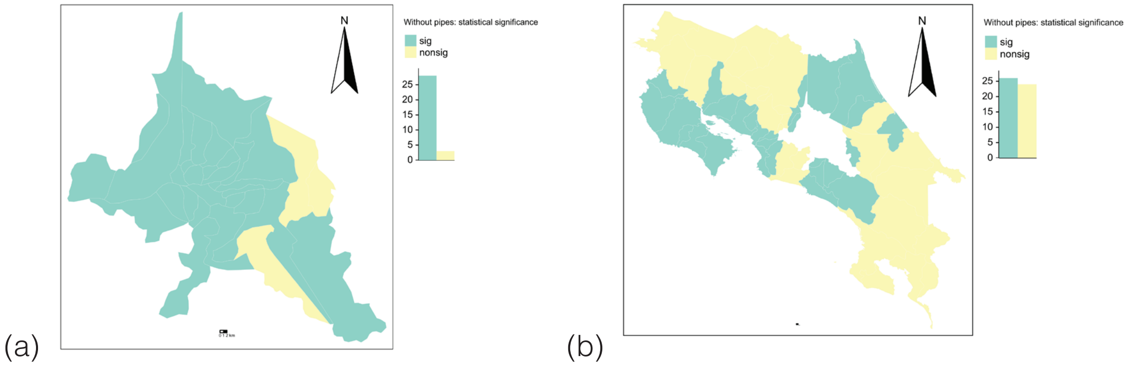

(a)
N
Without pipes: statistical significance
sig
nonsig
25
20
15
10
5
0
(b)
N
Without pipes: statistical significance
sig
nonsig
25
20
15
10
5
0

Figure 9. Statistical significance for population without water pipes at household (a) GAM (b) outside the GAM.

The statistical significance is evaluated for the results of the model. Figure 9 shows the case of the population without access to clean water, in the GAM region most of the regions show significance. Outside the GAM is almost the same significant and non-significant regions. The other variables can be found on the repository.

Summary

The results show that there are two possible relationships between socioeconomic variables and dengue cases that are worth exploring. The first is the employee mobility: there seems to be a positive correlation between minimum mobility (outside the labor force) and within the canton, while if they move outside the canton there seems to be a negative correlation. The behavior of the dengue mosquito is characterized by being a mosquito with little mobility, its life remains in the area near its breeding site. On the other hand, their feeding schedule is daytime. Considering these factors, it is relevant to analyze the mobility of people associated with their work, since work hours fit with the vector's feeding schedule.

The second is with the variable of dwellings with no access to water pipes, this variable presented a better fit in the model and shows a positive correlation with cases of dengue. We remark that there are more variables that influence this phenomenon of dengue, so these results cannot be interpreted as a direct relationship between these variables and cases of dengue, but they serve as a guide to identify the variables that can provide information that can be used in a more complex epidemiological model.

Conclusions and future work

The GWR model allows a geospatial analysis to evaluate if the variable that is being considered really has a possible relationship with the dependent variable, which helps to determine possible patterns in regions and a better understanding of the relationship between the variables related to dengue. It is an easy-to-implement model that allows a simple exploration of variables, in order to eventually determine whether or not to incorporate them into a more complex model.

The idea is to consider these variables to create an index that summarizes this socioeconomic information. This index is planned to be included as a parameter in an epidemiological model proposed by Sánchez and Calvo [13].

Access to data is usually challenging, not all variables are available for all the time ranges. Therefore, it would be interesting to have more detailed data for each canton updated by year, as well as data on the mobility of the different populations in the region, beyond the associated with the work.

In future iterations, the centroid of the regions could be better adjusted so that it is located in the most populated points of each canton, since it is currently only the centroid of the region, but the most populated area of the canton is not always located in the center.

Referencias

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] World Health Organization, UNICEF, et al. Operational guide using the web-based dashboard: Early warning and response system (ewars) fordengue outbreaks. 2020.

[3] Ministerio de Salud. Situación epidemiológica dengue, chikungunya y zika del ministerio de salud. data retrieved from: https://www.ministeriodesalud.go.cr/index. php/vigilancia-de-la-salud/analisis-de-situacion-de-salud. 2021

[4] Barrera R. Focks D. Dengue transmission dynamics: assessment and implications for control. In Report of the scientific working group meeting on dengue, 1-5, pages 92–108. WHO, October 2006.

[5] Rosen L. Rodhain F. Mosquito vectors and dengue virus-vector relation-ships. Dengue and dengue hemorrhagic fever, pages 45–60, 1997.

[6] Hwa-Lung Yu, Shang-Jen Yang, Hsin-Ju Yen, and George Christakos. A spatio-temporal climate-based model of early dengue fever warning in southern taiwan. Stochastic Environ- mental Research and Risk Assessment, 25(4):485–494, 2011.

[7] Christofer Åström, Joacim Rocklöv, Simon Hales, Andreas Béguin, Valerie Louis, and Rainer Sauerborn. Potential distribution of dengue fever under scenarios of climate change and economic development. Ecohealth, 9(4):448–454, 2012.

[8] Rachel Lowe, Bernard Cazelles, Richard Paul, and Xavier Rod ́o. Quantifying the added value of climate information in a spatio-temporal dengue model. Stochastic Environmental Research and Risk Assessment, 30(8):2067–2078, 2016.

[9] Elodie Descloux, Morgan Mangeas, Christophe Eugène Menkes, Matthieu Lengaigne, Anne Leroy, Temaui Tehei, Laurent Guillaumot, Magali Teurlai, Ann-Claire Gourinat, Justus Benzler, et al. Climate-based models for understanding and forecasting dengue epidemics. PLoS neglected tropical diseases, 6(2):e1470, 2012.

[10] Vivek Jason Jayaraj, Richard Avoi, Navindran Gopalakrishnan, Dhesi Baha Raja, and Yusri Umasa. Developing a dengue prediction model based on climate in tawau, malaysia. Acta tropica, 197:105055, 2019.

[11] M Hurtado-Díaz, H Riojas-Rodríguez, SJ Rothenberg, H Gomez-Dantés, and E Cifuentes. Impact of climate variability on the incidence of dengue in Mexico. Tropical medicine & international health, 12(11):1327–1337, 2007.

[12] Paola Vásquez, Antonio Loría, Fabio Sanchez, and Luis Alberto Barboza. Climate-driven statistical models as efective predictions of local dengue indicence in costa rica: a generalized additive model and random forest approach. Revista de Matemática: Teoría Y Aplicaciones, 27(1):1–21, 2020.

[13] Fabio Sanchez and Juan G Calvo. Dengue model with early-life stage of vectors and age- structure within host. Revista de Matemática: Teoría y Aplicaciones, 27(1):157–177, 2020.

[14] Cory W Morin, Andrew C Comrie, and Kacey Ernst. Climate and dengue transmission: evidence and implications. Environmental health perspectives, 121(11-12):1264–1272, 2013.

[15] Eric Delmelle, Michael Hagenlocher, Stefan Kienberger, and Irene Casas. A spatial model of socioeconomic and environmental determinants of dengue fever in cali, colombia. Acta tropica, 164:169–176, 2016.

[16] Chris Brunsdon, A. Stewart Fotheringham and Martin E. Charlton. Geographically weighted regression: a method for exploring spatial nonstationarity. Geographical analysis 28.4 (1996): 281-298.

[17] Naqvi, Syed Ali Asad, et al. ”Integrating Spatial Modelling and Space–Time Pattern Mining Analytics for Vector Disease-Related Health Perspectives: A Case of Dengue Fever in Pakistan.” International journal of environmental research and public health 18.22 (2021).

Improving Balanced Accuracy for Minority Plant Species under Data Imbalance

Mejorando la exactitud balanceada para especies de plantas minoritarias con datos desbalanceados

Ruben Gonzalez-Villanueva[1], Jose Carranza-Rojas[2]

Gonzalez-Villanueva, R; Carranza-Rojas, J. Improving balanced accuracy for minority plant species under data imbalance. *Tecnología en Marcha*. Vol. 37, special issue. August, 2024. IEEE International Conference on BioInspired Processing. Pag. 22-27.

https://doi.org/10.18845/tm.v37i7.7293

[1] Costa Rica Institute of Technology. Costa Rica.
rgonzalezv@estudiantec.cr
https://orcid.org/0000-0001-8044-3474

[2] Costa Rica Institute of Technology. Costa Rica.
jcarranza@itcr.ac.cr
https://orcid.org/0000-0002-9177-9173

Keywords

Imbalanced datasets;long-tail distribution;automatic plant identification;balanced metrics;deep learning; minority classes;classification.

Abstract

Regardless of the widely known success of deep learning in classification, such models are commonly measured by metrics that do not account for data imbalance, especially in terms of predictions per class, ignoring minority classes. This can be a problem, as minority classes are often the most difficult to predict and collect data for. In the plant domain, for example, species with fewer samples are often the ones that are hardest to collect and predict in the field. As we continue to identify more and more plant species, more of them become minority species, making it increasingly difficult to accurately classify them using traditional machine learning methods. To address this issue, we explore the combination of traditional data and machine learning approaches with deep learning techniques such as self-supervision in a preprocessing stage. By using self-supervised training together with different sampling algorithms and class weights, we were able to improve the balanced accuracy metric for minority plant species by between 7.9% and 13% without affecting general accuracy. This shows that using deep learning techniques in combination with traditional machine learning methods can help to improve the accuracy of predictions for minority classes, even in domains where data is limited.

Palabras clave

Conjuntos de datos desbalanceados; distribución de cola larga; identificación automática de plantas; métricas balanceadas; aprendizaje profundo; clases minoritarias; clasificación.

Resumen

A pesar del ampliamente conocido éxito del aprendizaje profundo en tareas de clasificación, estos modelos se miden comúnmente con métricas que no tienen en cuenta el desbalance de datos, especialmente en términos de predicciones por clase, ignorando las clases minoritarias. Esto puede ser un problema, ya que las clases minoritarias suelen ser las más difíciles de predecir y en términos de recolección de datos. En el dominio de las plantas, por ejemplo, las especies con un menor número de muestras son a menudo las más difíciles de recolectar y predecir en el campo. A medida que se siguen identificando más y más especies de plantas, más de ellas se vuelven minoritarias, lo que dificulta cada vez más la clasificación precisa utilizando métodos tradicionales de aprendizaje automático. Para abordar este problema, se explora la combinación de enfoques de los datos y tradicionales de aprendizaje automático con técnicas de aprendizaje profundo, como la auto-supervisión en una etapa de preprocesamiento. Al utilizar el entrenamiento auto supervisado junto con diferentes algoritmos de muestreo y pesos de clase, logramos mejorar la métrica de exactitud balanceada para las especies de plantas minoritarias entre el 7.9% y el 13% sin afectar la datos general. Esto demuestra que el uso de técnicas de aprendizaje profundo en combinación con métodos tradicionales de aprendizaje automático puede ayudar a mejorar la precisión de las predicciones para clases minoritarias, incluso en dominios donde los datos son limitados.

Introduction

Imbalanced datasets affect models to classify species more fairly [1]. Training a classifier with a long-tailed distribution may achieve good results on the general accuracy metric, but not on a metric that considers the predictive ability of each class, even minority ones, as the balanced accuracy metric does [2].

Different approaches have been tried to solve this problem and find a configuration to more fairly classify an imbalanced dataset, but these approaches only use traditional ML methods [3],[4]. In this short paper we use these ML approaches, combine them with DL approaches such as self-supervision pre-training, to get better results on the balanced accuracy metric for minority species without decreasing overall accuracy in a plants dataset.

Matherials and methods

Materials

Dataset

The CRLeaves dataset contains images of leaves from Costa Rica from 255 species. It was taken from [5]. It contains a total of 6,938 images but has a long-tail distribution in which the species with the most images contain 89, while the species with the fewer images contains only 4. All images have an uniform background as shown in the figure 1.

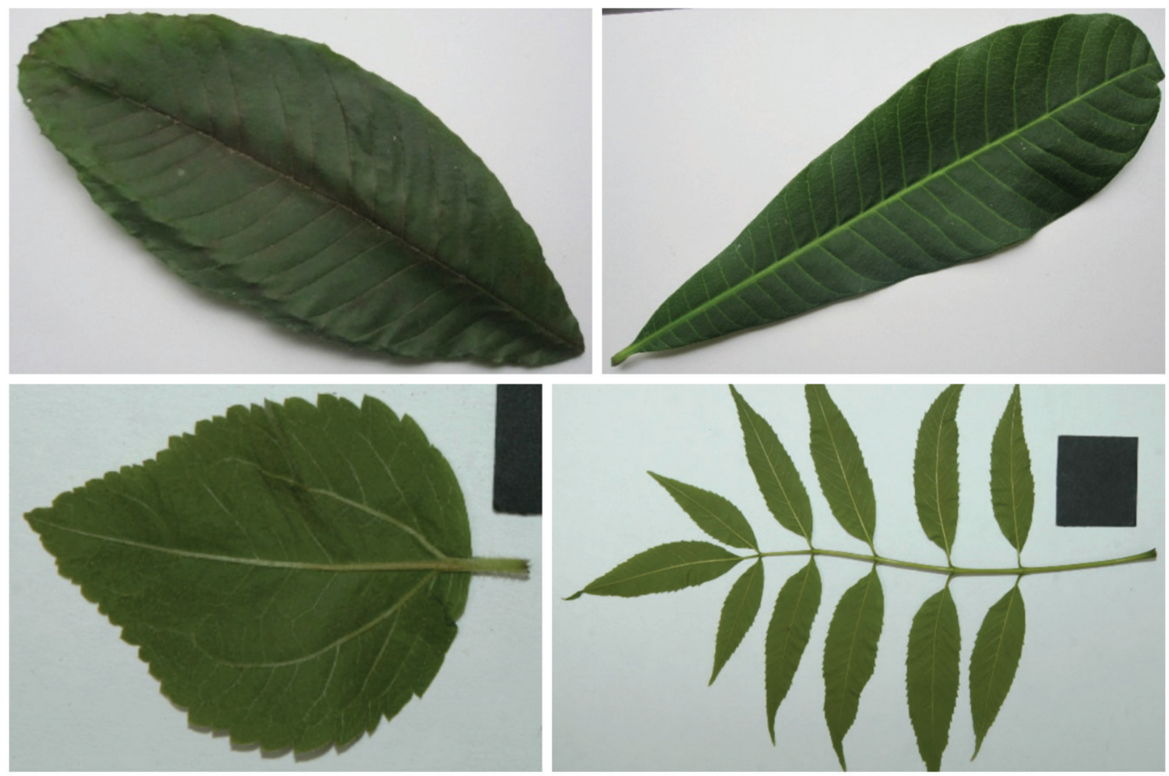


Figure 1. Samples from CR Leaves dataset of Clethra costaricensis, Anacardium excelsum, Calea urticifolia and Tecoma Stans. Source: [5]

Architecture

The chosen deep learning architecture is ResNet50 [6]. It contains 50 deep residuals layers and keeps good results in computer vision for our dataset. This is the baseline for the experiments.

Metrics

We used the general accuracy that calculates the hit rate of the predictions and labels for all the samples [2]. Additionally, we focus on balanced accuracy that uses weights in the general accuracy to measure the predictions by class [7]. Balanced accuracy for minority species is calculated only for the species with less than 10 samples.

Methods

Pre-Training Methods

The chosen pre-training method is SimCLR [8]. It is a self-supervision approach of encoder-decoder used to pre-train the model with no labels. SimCLR contains two parts: training method and the architecture change that consists in a new fully connected (FC) layer at the end of the model. We test how both training methods work with the baseline architecture (ResNet50) and with the SimCLR FC.

Imbalanced Algorithms

The chosen algorithms to address the imbalance problem are from two different kind of approaches. The first one is the data approach in which we used random undersampling and random oversampling to obtain a similar number of images per species [3]. We explored two implementations:

Sampling A

We use the implementation of sklearn [9]. It adjusts the weights inversely proportional to class frequencies as shown in Equation 1, where is the weight for class , *S* the number of samples and *C* the number of classes.

$$w_i = \frac{S}{C \, * \, S_i} \tag{1}$$

Sampling B

It is described in Equation 2. It divides a 1 with the number of samples in the class to obtain the inverse frequency. We use NumPy's implementation.

$$w_i = \frac{1}{S_i} \tag{2}$$

Class Weights

We additionally use class weights, which are based on calculating the weights for each species and give this information to the cross-entropy loss [4].

Results

Table 1 describes the factors for the experiments, which combined sum up to 12 experiment runs. We ran 3 repetitions for each combination for statistical validity. We ran the SimCLR self-supervision pre-training models for 100 epochs to use the weights in the experiments. For the supervised training, we used a distribution of 50:50 for training and testing.

Table 1. Factors for experiments.

Pre-Training Methods	Imbalanced Algorithms
None	None
SimCLR Training	Sampling A
Sim CLR Training + FC	Sampling B
	Class Weights

We report the P-Value from the pairwise T test, as well as the results for the metric General Accuracy in Table 2. The baseline model without any techniques for imbalance data obtained 0.87 of general accuracy. The only combination that was statistically different from the baseline was SimCLR Training + FC + class weights with an improvement. In the other combinations there was no statistical difference. This means the combinations of algorithms to address the data imbalance are not hurting drastically the general accuracy.

In contrast, the results for Balanced Accuracy Minority Species are also described in Table 2. The baseline obtained a value of 0.657, been this a difference of 0.213 compared to general accuracy. There were 3 combinations that were statistical better than the baseline: Sampling A, Sampling B and Class Weights, all combined with SimCLR Training.

Table 2. Average balanced accuracy for minority species, average general accuracy, and p-values for interactions of factors vs the baseline with the testing subset. ↑ denotes statistical improvement against the baseline.

Pre-Training Method	Imbalanced Algorithm	Average Balanced Accuracy Minority Species / P-Value	Average General Accuracy / P-Value
None	None	0.657	0.870
	Sampling A	0.713 / 0.390 (-)	0.865 / 0.580 (-)
	Sampling B	0.713 / 0.270 (-)	0.867 / 0.640 (-)
	Class Weights	0.722 / 0.180 (-)	0.869 / 0.890 (-)
SimCLR Training	None	0.694 / 0.270 (-)	0.883 / 0.068 (-)
	Sampling A	0.787 / 0.020 (↑)	0.881 / 0.190 (-)
	Sampling B	0.736 / 0.023 (↑)	0.877 / 0.260 (-)
	Class Weights	0.750 / 0.042 (↑)	0.880 / 0.160 (-)
Sim CLR Training + FC	None	0.690 / 0.500 (-)	0.895 / 0.097 (-)
	Sampling A	0.727 / 0.110 (-)	0.885 / 0.11 (-)
	Sampling B	0.755 / 0.120 (-)	0.886 / 0.042 (↑)
	Class Weights	0.713 / 0.210 (-)	0.886 / 0.051 (-)

Conclusions and recommendations

The results show that there was only one experiment that had an improvement in general accuracy, but this combination did not improve the balanced accuracy for minority species. For the balanced accuracy for minority species metric, there was no improvement using pre-training methods or the imbalanced algorithms individually. We found three combinations that had an improvement on imbalanced accuracy, in particular when there was a combination of SimCLR Training as Pre-Training Method and with sampling methods and class weights. This shows that it is possible to improve imbalanced metrics without heavily affecting balanced metrics.

Finally, for this dataset the best way we found to improve the metric is to combine SimCLR Training and the different imbalanced algorithms used. Future work includes performing these experiments on a larger dataset with more species, images, and even more imbalance. In addition, finding completely new approaches to further improve the results obtained, which could be based on loss functions using the plant hierarchy.

Acknowledgments

We would like to thank the Costa Rica Institute of Technology for its support during this research.

References

[1] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, “A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 4, pp. 463–484, 2012.

[2] N. Bressler, “How to check the accuracy of your machine learning model,” Feb 2022. [Online]. Available: https://deepchecks.com/how-to- check-the-accuracy-of-your-machine-learning-model/

[3] Y. Pristyanto, I. Pratama, and A. F. Nugraha, “Data level approach for imbalanced class handling on educational data mining multiclass classification,” in 2018 International Conference on Information and Communications Technology (ICOIACT), 2018, pp. 310–314.

[4] S. Lu, F. Gao, C. Piao, and Y. Ma, “Dynamic weighted cross entropy for semantic segmentation with extremely imbalanced data,” in 2019 Interna- tional Conference on Artificial Intelligence and Advanced Manufacturing (AIAM), 2019, pp. 230–233.

[5] J. Carranza-Rojas and E. Mata-Montero, “Combining leaf shape and texture for costa rican plant species identification,” CLEI Electronic journal, vol. 19, no. 1, pp. 7–7, 2016.

[6] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[7] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, “The balanced accuracy and its posterior distribution,” in 2010 20th International Conference on Pattern Recognition, 2010, pp. 3121–3124.

[8] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple frame- work for contrastive learning of visual representations,” in International conference on machine learning. PMLR, 2020, pp. 1597–1607.

[9] G. King and L. Zeng, “Logistic regression in rare events data,” Political analysis, vol. 9, no. 2, pp. 137–163, 2001.

Wearable device for detecting flat feet and high arches using pressure sensors based on graphite

Dispositivo vestible para detectar pie plano y cavo utilizando sensores de presión a base de grafito

Tatiana Dolores Cárdenas Guaraca[1], Kely Thalía Aucaquizhpi Inga[2], Nimrod Isaias Sarmiento Salamea[3], Katherine Yomara Berrezueta Barrezueta[4], Angel Andres Perez Muñoz[5]

Cárdenas-Guaraca, T. D; Aucaquizhpi-Inga, K. T; Sarmiento-Salamea, N. I; Berrezueta-Barrezueta, K. Y; Perez-Muñoz, A. A. Wearable device for detecting flat feet and high arches using pressure sensors based on graphite. *Tecnología en Marcha*. Vol. 37, special issue. August, 2024. IEEE International Conference on BioInspired Processing. Pag. 28-37.

https://doi.org/10.18845/tm.v37i7.7294

[1] GI-IATa, UNESCO Chair in Assistive Technologies for Educational Inclusion, Universidad Politécnica Salesiana, Cuenca, Ecuador.
tcardenas@est.ups.edu.ec
https://orcid.org/0009-0009-1738-1990

[2] Universidad Politécnica Salesiana. Ecuador.
kaucaquizpi@est.ups.edu.ec
https://orcid.org/0009-0005-2704-7890

[3] Universidad Politécnica Salesiana. Ecuador.
nsarmientos@est.ups.edu.ec
https://orcid.org/0009-0004-7881-8811

[4] Universidad Politécnica Salesiana. Ecuador.
kberrezuetab@est.ups.edu.ec
https://orcid.org/0009-0001-0432-3128

[5] GI-IATa, UNESCO Chair in Assistive Technologies for Educational Inclusion, Universidad Politécnica Salesiana, Cuenca, Ecuador.
aperezm@ups.edu.ec
https://orcid.org/0000-0003-4896-723X

Keywords

Medical technology; remote sensing; electronic engineering; biomedical research; device; plantar conditions; Android application.

Abstract

The present study focuses on the design and construction of an innovative device aimed at identifying plantar foot conditions, such as normal, cavus, and flat feet, through the analysis of biosignals. The device was based on two identification methods: the plantar footprint test and the use of a template equipped with digital and analog sensors made of Velostat paper, a material derived from graphite that allows capturing biosignals with high precision. These biosignals were integrated into an Android application, facilitating the diagnosis and appropriate treatment of plantar conditions. The key component of the device was a sensor created with graphite, which functioned as a piezoelectric sensor to acquire and measure the pressure exerted by the foot. Thanks to the sensor's ability to capture biosignals, it was possible to accurately detect and classify plantar conditions. The device stood out for its low cost and seamless integration with the mobile application, becoming a valuable and accessible tool for the healthcare sector, with the potential to significantly improve the quality of life for patients with foot problems through early detection and timely treatment.

Palabras clave

Tecnología médica; detector; tecnología electrónica; investigación médica; dispositivo; condiciones plantares; aplicación de Android.

Resumen

El presente estudio se centra en el diseño y construcción de un dispositivo innovador destinado a la identificación de condiciones plantares del pie, como el pie normal, cavo y plano, mediante el análisis de bioseñales. El dispositivo se basó en dos métodos de identificación: la prueba de la huella plantar y el uso de una plantilla equipada con sensores digitales y analógicos fabricados con papel Velostat, un material derivado del grafito que permite capturar bioseñales con alta precisión. Estas bioseñales se integraron en una aplicación de Android, lo que facilita el diagnóstico y tratamiento adecuado de las condiciones plantares. El componente clave del dispositivo fue un sensor creado con grafito, que operó como un sensor piezoeléctrico para adquirir y medir la presión ejercida por el pie. Gracias a la capacidad de este sensor para captar bioseñales, fue posible detectar y clasificar con precisión las condiciones plantares. El dispositivo se destacó por su bajo costo y su perfecta integración con la aplicación móvil, convirtiéndose en una herramienta valiosa y accesible para el sector de la salud, con el potencial de mejorar significativamente la calidad de vida de los pacientes con problemas en los pies mediante la detección temprana y el tratamiento oportuno.

Introduction

This project focuses on the development of an insole for the detection of various plantar foot conditions, such as normal foot, four degrees of flat foot and three degrees of pes cavus. In addition, temperature measurements are taken to determine whether the person being tested is suffering from diabetes, using temperature sensors built into the insole. If the person is diabetic, a temperature difference of 2.2°C at the same point between the two feet will be observed, which could also indicate the possibility of ulcers. A difference of 2.2°C at the same point on both feet, for a given individual, suggests the appearance of characteristic diabetic foot lesions, such as ulcers or Charcot arthropathy. Scientific studies support the idea that foot temperature control reduces the incidence of ulcers in Diabetes Mellitus [1]. It should be noted that the insole has a relevant feature: a body mass limit of 75 kilograms (kg). Posture, in general, depends on several factors, including the presence of normal arches in the foot, vertical alignment of the ankles and a balanced distribution of body mass around the center of gravity. The foot, as a specialized structure, allows for proper distribution of static and dynamic loads [2].

The main objective of this project is to obtain a portable and low-cost solution for the detection of these conditions, since it has been found that the podoscopes available on the market are expensive and their accessibility is difficult for those who need them. Although there are podoscopes with excellent performance, their price is excessively high, which is why numerous investigations have been carried out to design and manufacture affordable podoscopes that offer reliable results [3]

Materials and methods

Materials

The materials used in this project are as follows:

- Small protoboard
- 2 resistors of 220 Ω
- Insulating tape
- 5 large jumpers of different colors
- 5 large red jumpers
- 5 female-female or male-female jumpers
- DHT11 temperature sensor
- Esp32
- Velostat
- Double-sided tape
- 3 push-button modules
- Mat
- Size 38 of shoe insole
- Flex foam

Methods

The method used in this study is known as the "Plantar Footprint Method". This approach involves assessing the shape and functionality of the foot, diagnosing foot-related problems, and determining the need for custom orthotics and footwear. However, due to its high cost, not all clinics have the necessary equipment, which can lead to medical malpractice [4].

Research on plantar conditions

A normal foot has a specific ratio of A-A' to B-B' measurements in a footprint, but flat and cavus feet have abnormal ratios. However, slight variations may not be pathological [5].

Flatfoot is an inherited condition that affects the inner arch of the foot. In children under three years old, it can be easily detected through an examination using a podoscope. The flatfoot footprint has a flat inner curve [6].

Hollow or high-arched feet have a A normal foot is characterized by a specific relationship between the A-A' and B-B' measurements in a plantar footprint, whereas flat and cavus feet show abnormal relationships. However, it is important to note that slight variations may not necessarily be pathological [5].

Flatfoot is an inherited condition that affects the internal arch of the foot. In children under three years of age, this condition can be easily identified by examination using a podoscope. In the plantar footprint of flat feet, a flat inner curve is observed [6].

On the other hand, hollow feet, also known as high-arched feet, are characterized by an abnormally high arch, which can lead to instability in walking and health problems, such as sprains, overloads and muscle dysfunction. It is recommended that children's feet be examined from the age of 3 years to verify their correct development, since hollow feet can lead to the formation of calluses and metatarsalgia [7].

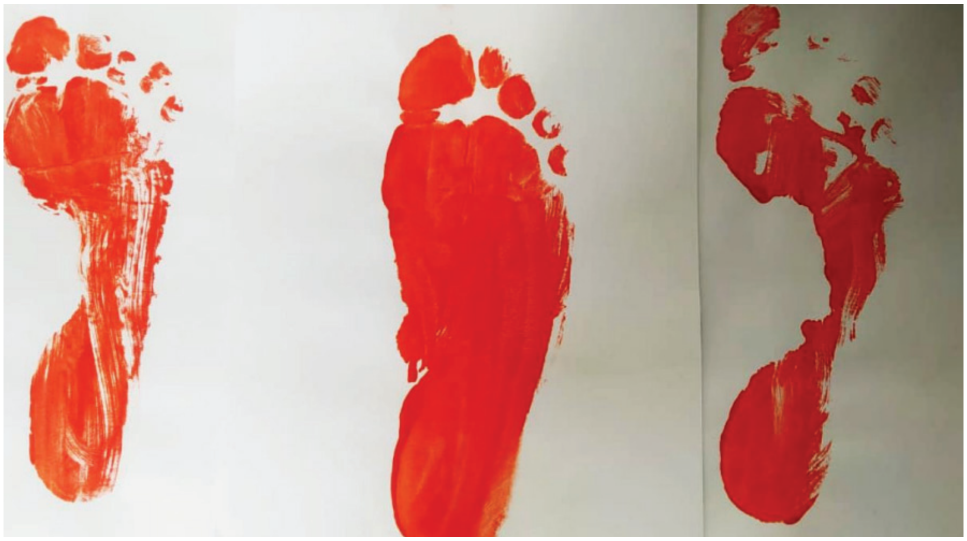


Figure 1. Plantar condition tests

Data to consider

To determine the maximum load that the sensor can withstand, the weight distribution on both legs of the subject and on the five sensors of the device was considered. Let us assume that the person has a weight of 50 kg, which is equivalent to 490 N, considering the acceleration of gravity of 9.8 m/s^2.

Performing the load distribution, an approximate force of 50 N is obtained. The area of the sensor, which corresponds to a circle with a diameter of 2 cm, is calculated using the formula for the area of a circle:

$$A = \pi \times r^2$$

Substituting the values, we obtain:

$$A = \pi \times (0.01m)^2$$
$$A = 3.1416 \times 10^{-4}\ m^2$$

With the area of the device, we can calculate the pascals applied with the weight of 50 N.

$$P = \frac{N}{A}$$
$$P = \frac{50\ N}{3.1416 \times 10^{-4}\ m^2}$$
$$P = 159154.9431\ Pa$$
$$P = 0.15992\ MPa$$

With this result, mechanical simulations were carried out to evaluate the behavior of the sensor. Similarly, the calculation was performed for a force of 100 N.

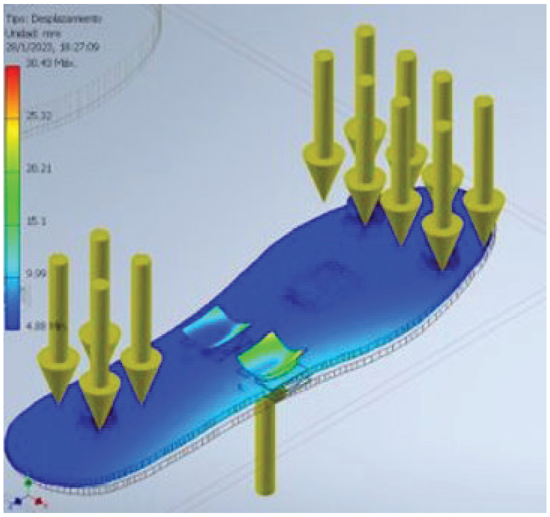


Figure 2. Simulation for 50 N.

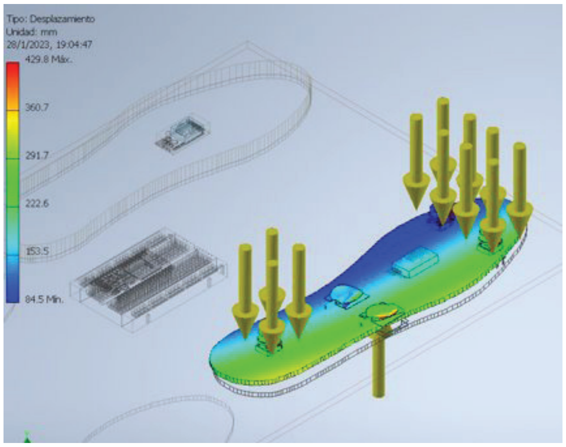


Figure 3. Simulation for 100 N

Results

In the study, Linqstat, a conductive material that has the ability to create flexible sensors by changing its resistance upon application of force, was used. In addition, Velostat, a polymer film with piezoresistive and electrically conductive properties, was used [8].

The combination of these materials allowed the development of pressure sensors with a simple structure consisting of three layers: two conductive wires and an insulator.

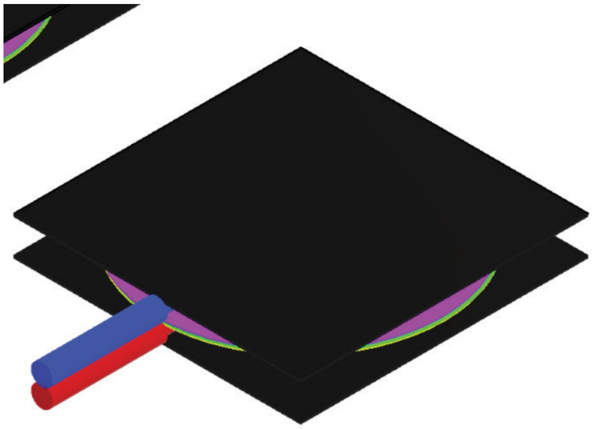


Figure 4. Pressure sensor

To carry out the characterization of the sensor, we proceeded to gradually apply weights on it, taking into account their respective values. The voltage change corresponding to each weight applied was recorded, making the records every 15 seconds to allow the stabilization of the sensor and obtain more accurate results.

Figure 5 shows the calibration curve of the pressure sensor, which was obtained through three tests using the aforementioned methodology and calculating the average of the results obtained. The curve provides information on the relationship between the recorded voltage and the weights applied to the sensor. In addition, the equation of the curve was used to obtain the specific characteristics of the sensor, taking as midpoint the value of 49.26 on the y-axis.

A detailed summary of the characteristics of the manufactured pressure sensor derived from the calibration curve is presented in Table 1. The table provides relevant information on the detectable pressure range, sensor sensitivity and other parameters important for understanding and effectively using the sensor in specific applications.

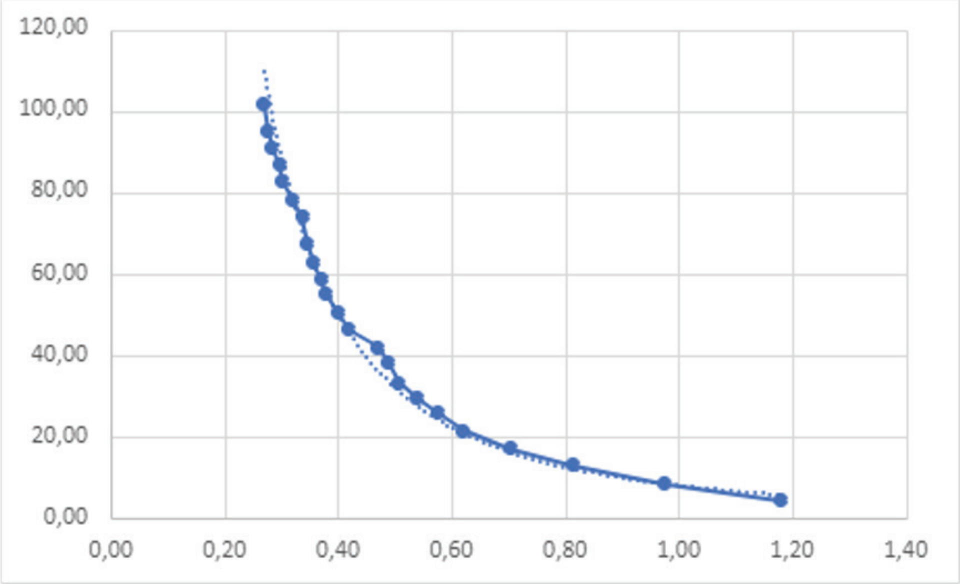


Figure 5. Calibration curve of the pressure sensor.

$$y = 8.0754x^{-1.994} \tag{1}$$

$$49.26 = 8.0754x^{-1.994} \tag{2}$$

$$x = \sqrt[1.994]{\frac{8.0754}{49.26}} \tag{3}$$

$$x = \pm 0.403$$

$$Accuracy\ Error: \frac{100 - 0.403}{100} = 0.995\% \tag{4}$$

$$Repeatability\ Error: \frac{0.18}{101.61} \times 100 = 0.177\% \tag{5}$$

Table 1. Pressure Sensor Features.

Features	Value
Calibration curve	y= 8.0754x-$^{1.994}$
Sensitivity	8.0755 $\frac{N}{V}$
Offset	*2,84 V*
Repeatability Error	0.17%
Accuracy Error	0.995 %

Analysis of the sensor characteristics reveals that participants were able to obtain accurate results about the condition of their feet using the developed system, with an average error of 0.995%. In addition, tests performed to obtain the plantar footprint confirmed the ability of the device to identify the state of the indicated foot.

In the first prototype of the device, as shown in Figures 6 and 7, the electronics were housed in a 3D printed box, while the sensors were strategically placed inside insoles for easy access and use by users.

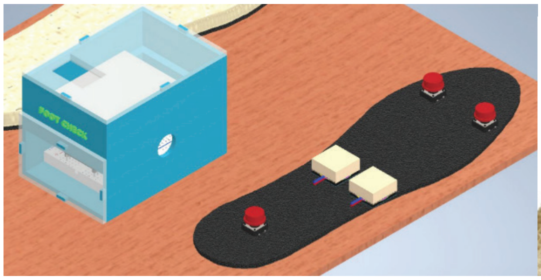


Figure 6. Device modeling.

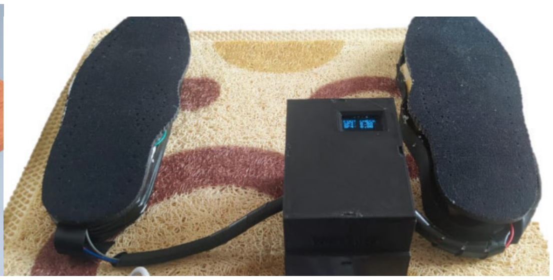
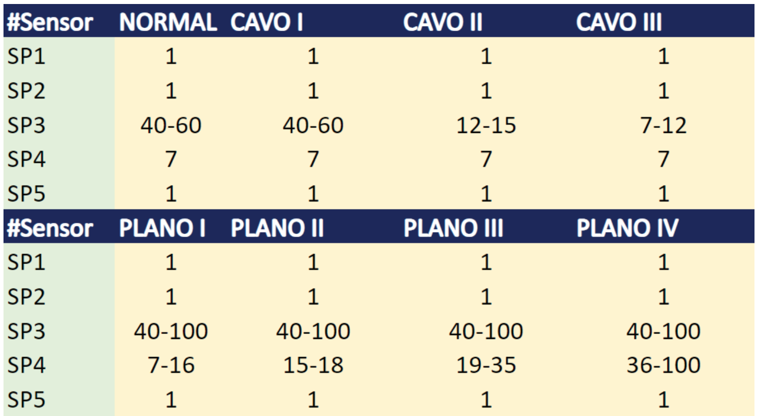


Fig. 7. First prototype.

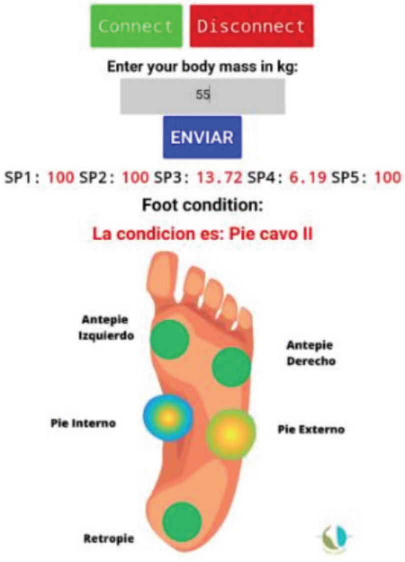
The conditions for the sensor were established using pressure percentages adjusted through testing with multiple subjects. These conditions were integrated into the system programming. Figure 8 shows the conditions based on the defined pressure percentages. For the control of the ESP32 microcontroller, Arduino programming was used, which includes the characterization of the sensor equation and all necessary commands.



#Sensor	NORMAL	CAVO I	CAVO II	CAVO III
SP1	1	1	1	1
SP2	1	1	1	1
SP3	40-60	40-60	12-15	7-12
SP4	7	7	7	7
SP5	1	1	1	1
#Sensor	**PLANO I**	**PLANO II**	**PLANO III**	**PLANO IV**
SP1	1	1	1	1
SP2	1	1	1	1
SP3	40-100	40-100	40-100	40-100
SP4	7-16	15-18	19-35	36-100
SP5	1	1	1	1

Figure 8. Condiciones del pie en base a porcentajes de presión.

In order to visualize and obtain real-time data, a user interface was developed (Figure 9). This interface allows user interaction and real-time monitoring of the patient's plantar condition. The application is easy to use and provides essential information, such as company details, different plantar conditions available, and the sensors being monitored.



Enter your body mass in kg: 55

SP1: 100 SP2: 100 SP3: 13.72 SP4: 6.19 SP5: 100

Foot condition:

La condicion es: Pie cavo II

Figure 9. App interface.

The app's interface shows a visual representation of the feet, where three support points are identified by digital sensors in green, and two analog sensors that indicate different levels of plantar involvement. These sensors light up in proportion to the percentage of pressure applied.

Conclusions

Currently, the "FOOT CHECK" device is designed for real-time monitoring, however, work is underway to improve its structure to allow users to use it for longer periods of time. The main objective is to store the data obtained to facilitate more effective patient follow-up in the future. Since plantar conditions are more susceptible to correction in children, the development team is exploring ways to make the biomedical images interactive in the app, in order to increase patient engagement in the use of the device.

The "FOOT CHECK" device along with its app, offers an affordable and accessible alternative to the podoscope for diagnosing plantar conditions. Although still under development, the device is already functional and connects to the ESP32 server via Bluetooth to determine and monitor plantar conditions. To ensure proper use of the device, manuals have been developed to provide the necessary instructions.

Acknowledgments

We sincerely appreciate the support received to carry out this project in the Biomedical Engineering program at the Universidad Politécnica Salesiana, as well as the valuable opportunity to present our work at the EDS-IEEE event. We want to express our special thanks to the Artificial Intelligence Research Group (GIIATA) of the university and the event organization for their continuous support. Our team is excited and committed to continue advancing in our research, focused on the development of innovative sensors and devices, with the aim of contributing to the progress of biomedicine and health in general. We are dedicated to finding technological solutions that improve people's quality of life and provide significant advancements in the field of medicine.

Reference

[1] D. Simba and M. Tipán, “repositorio.puce.edu.ec,” Jan. 2018. [Online]. Available: http://repositorio.puce.edu.ec/bitstream/handle/22000/14742/TESIS%20DANIELA%20S.%20%26%20MAYRA%20T..pdf?sequence=1&isAllowed=y. [Accessed: Jul. 15, 2023].

[2] De La Peña, R. G., Benhamú, S. B., Cristino, M. D. J., Del Arco, J. G., & Noguerón, G. A. G. (2019). La temperatura del pie como factor predictivo de aparición de úlceras en la diabetes mellitus. Revista internacional de ciencias podológicas, 13(2), 115-129. https://doi.org/10.5209/ricp.64726

[3] N. C. C. Morales, “repositorio.ucundinamarca.edu.co,” Feb. 28, 2019. [Online]. Available: https://repositorio.ucundinamarca.edu.co/bitstream/handle/20.500.12558/2372/DESARROLLO%20DE%20UN%20SISTEMA%20PARA%20LA%20IDENTIFICACI%c3%93N%20DE%20ALTERACIONES%20EN%20LA%20POSTURA%20MEDIANTE%20EL%20AN%c3%81LISIS%20DE%20LA%20HUELLA%20PLANTAR.pdf?sequence=. [Accessed: Jul. 15, 2023].

[4] “Ángulos pie - centre podomèdic,” Centre Podomèdic, Jun. 28, 2020.

[5] M, L. P. (2003, 1 noviembre). Alteraciones de la bóveda plantar. Revista Española de Reumatología. https://www.elsevier.es/es-revista-revista-espanolareumatologia-29-articulo-alteraciones-bovedaplantar-13055069 .

[6] http://themeforest.net/user/dan_fisher. (s. f.). Ortopédicos Dinky. https://www.ortopedicosdinky.com.mx/pato-infantil.html

[7] “Pie Cavo,” Centre Podomèdic. [Online]. Available: https://www.centrepodomedic.com/blog/angulospie.

[8] “Hoja conductiva tipo Velostat,” Electrónica Steren México. [Online]. Available: https://www.steren.com.ec/hojaconductiva-tipo-velostat.html. [Accessed: Jan. 22, 2023

Proposal of self and semi-supervised learning for imbalanced classification of coronary heart disease tabular data

Propuesta de aprendizaje auto-semi supervisado para la clasificación desequilibrada de datos tabulares de enfermedades coronarias

Danny Xie-Li[1], Manfred González-Hernández[2]

Xie-Li, D; González-Hernández, M. Proposal of self and semi-supervised learning for imbalanced classification of coronary heart disease tabular data. *Tecnología en Marcha*. Vol. 37, special issue. August, 2024. IEEE International Conference on BioInspired Processing. Pag. 38-43.

https://doi.org/10.18845/tm.v37i7.7295

[1] Instituto Tecnológico de Costa Rica. Costa Rica.
dxie@ic-itcr.ac.cr
https://orcid.org/0000-0003-1878-9460

[2] Universidad de Costa Rica. Costa Rica.
manfred.gonzalezhernandez@ucr.ac.cr
https://orcid.org/0000-0002-5408-7901

Keywords

Self-supervised learning; semi-supervised learning; data augmentation; contrastive learning; imbalanced; medical datasets.

Abstract

Triple Mixup is an augmentation policy in the hidden latent space we introduced in the Contrastive Mixup Self-Semi Supervised learning framework, to address the imbalanced data problem, for Cardiovascular Heart Diseases tabular dataset. Medical tabular datasets are known to present challenges as high imbalanced class, limited annotated quality samples due to the domain nature. Recent literature in Self and Semi supervised learning, has shown tremendous progress in learning useful representations, and leveraging unlabeled dataset and labeled dataset to train a learning model. Most existing methods are not feasible for tabular data due to the data augmentation scheme. In addition, the high imbalanced problem can show lower performance on machine learning algorithms. For this work, we propose the triple data augmentation method in hidden space to attack the unbalanced challenge in self-supervised and semi-supervised learning, from the possible applications of Contrastive Mixup, thus we will study the influence of it.

Palabras clave

Aprendizaje Autosupervisado, Aprendizaje Semisupervisado, Aumentación de datos, Aprendizaje por Contraste, Desbalance de datos, datos médicos.

Resumen

Triple Mixup es una política de aumento en el espacio latente oculto que introdujimos en el marco de aprendizaje autosupervisado de Mixup contrastivo, para abordar el problema de datos desequilibrados, para el conjunto de datos tabulares de enfermedades cardíacas cardiovasculares. Se sabe que los conjuntos de datos tabulares médicos presentan desafíos como muestras de calidad anotada limitada y de clase altamente desequilibrada debido a la naturaleza del dominio. La literatura reciente sobre el aprendizaje autosupervisado y semi supervisado ha mostrado un enorme progreso en el aprendizaje de representaciones útiles y en el aprovechamiento de conjuntos de datos no etiquetados y conjuntos de datos etiquetados para entrenar un modelo de aprendizaje. La mayoría de los métodos existentes no son factibles para datos tabulares debido al esquema de aumento de datos. Además, el problema de alto desequilibrio puede mostrar un rendimiento más bajo en los algoritmos de aprendizaje automático. Para este trabajo, proponemos el método de aumentación de datos triple en el espacio oculto para atacar el desafío desequilibrado en el aprendizaje autosupervisado y semi supervisado, desde las posibles aplicaciones de Contrastive Mixup, por ende estudiaremos la influencia de este .

Introducción

Diversity on the training data is a key piece in the process of generalization, when training a supervised machine learning model. However, this can become a limitation due to the number and quality of the ground truth or related to how well diverse is the data. This problem has been attacked by data augmentation techniques in computer vision approaches [1]– [5] and, recent work has extended to tabular data approaches [6], [7]. However, tabular data often contain heterogeneous features that represent a mixture of continuous, categorical and ordinal values [8], and there is not an inherent positional information.

Unfortunately, obtaining labeled data is often infeasible in the healthcare domain, as annotation requires domain expert and manual labor. In addition, concerned with a particularly low representation of classes such as rare diseases. However, is often a wealth of unlabeled data available, but annotated data is only available for a small group. In order to take advantage of the unlabeled data, semi-supervised learning leverages the use of labeled and unlabeled data on training. Existing semi-supervised learning (SSL) algorithms from image and text domains are not effective for tabular data, because they heavily rely on spatial or semantic structure [9].

Some Tabular Health datasets are highly imbalanced, are often one or some minority classes, and most of the cases, the ”minority” are more important than the ”major” classes. As a result, traditional machine learning methods tend to produce results overwhelmed by the majority of classes, decreasing the model prediction accuracy. However, SSL has a limited assumption that the number of samples in different classes are balanced, and show lower performance for imbalanced class distribution datasets [10]. To augment data, the authors [11] explore the use of MixUp in the hidden space as augmentation for tabular data, and to consider class imbalanced [12] triple mixup to augment the data to generate more minority examples, but is only considered in the early stage for continuous variable.

Self-supervised learning frameworks have proven to be effective to learn representations from unlabeled data [13]. It is capable of adopting pseudo-labels based on the attributes learned for several downstream prediction tasks [11]. In addition, many self-supervised methods are based on contrastive representation learning. The authors in [14] for visual representation, extend the use of augmentation to generate ”similar” samples, and normalized representations based on contrastive cross-entropy loss, to minimize the distance in the latent space of ”positive” pairs and maximizes the distance of ”negative” pairs. Recent approaches extend the use of self-supervised on tabular domain, authors in [9], proposed to recover the mask vector, in addition to the original sample with a novel corrupted sample generation for feature representation [11].

As indicated by [15], Cardiovascular Heart Diseases (CVDs) remain as one of the death causes with the highest mortality rates in the world. An estimated 17.9 million lives taken per year as mentioned by the World Health Organization. Although, several approaches focused on machine learning algorithms for CVDs have been proposed in the early years. This is the case for Chronic Kidney Disease where [16] suggests the implementation of Random Forest [17], a Supervised Machine Learning (SML) model to predict the occurrence of the disease. [18] proposed SML models to deal with the Coronary Heart Disease using the standard data set [19]. They opted to duplicate rows and get equal quantities for the number of rows they had per each class to manage the unbalanced data in the data set. Random Forest, Decision Tree and K-nearest neighbors were used in [18] to predict the existence of the deceased. In our proposed method we deal with the unbalanced data set [19] -collected by The Framingham Township Heart Institute offering a 10-year data set on coronary heart disease-via Triplet Mix up and a semi-supervised machine learning model.

Preliminaries

To present our proposed method, we extend the work of [11] and [12] to formally introduce self-supervised, contrastive loss and semi-supervised loss. Given a dataset with *N* examples, there is a small subset defined as $D_L = (x_i, y_i)^{N_L}{}_{i=1}$ for each example with the corresponding label, and $D_U = (x_i, y_i)^{N_L}{}_{i=1}$ defined as the unlabeled dataset. Consider the *x* as the input features (consisting of numerical and categorical features) where $y_i \in \{0,1\}, 1$ indicates that the person is prone to suffer CHD and 0 indicates the person is less likely to suffer CHD. In downstream tasks, we use supervised learning to find an optimized function $f(X)$ that $f(X) \rightarrow Y$ to minimize given the loss function *l*.

Self-Supervised Learning: Contrastive Loss

Inspired by the recent contrastive learning framework [20] using metric learning methods, to learn representations. Given a batch of N samples is augmented using an augmentation function *Aug(.)* to create a multi-viewed batch with *2N* pairs, $\{\tilde{x}_i, \tilde{y}_i\}_i = 1, \ldots, 2N$ where $\tilde{x}_{2k}$ and $\tilde{x}_{2k-1}$ are two random augmentations of the same sample x_k for $k = 1, \ldots, N$. The samples are fed to an encoder $e : x \rightarrow z$, which takes a sample $x \in X$ to obtain a latent representation $z = e(x)$. For the pretext task defined, the model is trained jointly to minimize the self supervised contrastive loss function *l*.

$$\underset{\square}{min}\, e, h\ \mathbb{E}_{(x,y)} \sim P(X, \tilde{Y})[l(\tilde{y},\ h(e(x)))] \tag{1}$$

Where *h* maps *z* to an embedding space $h : z \rightarrow v$. Within a multi viewed batch $i \in I = \{1, \ldots, 2N\}$ the self supervised contrastive loss is defined as

$$l = \sum_{i \in I}^{\square} \square -log\ [\frac{exp(sim(v_i, v_j(i))/\tau)}{\sum_{n \in I \setminus \{i\}}^{\square} \square\ exp(sim(v_i, v_n)/\tau)}] \tag{2}$$

where $sim(.,.)$ ∈ R+ is a similarity function (e.g., dot product or cosine similarity), τ ∈ R+ is a scalar of temperature parameter, *i* is the anchor, *A(i)* is the positive(s) and $I(i)$ are the negatives. The positive and negative samples refer to samples that are semantically similar and dissimilar, respectively. Intuitively, the objective of this function is to bring the positives and the anchor closer in the embedding space v and opposite for the negative samples.

Semi-Supervised Learning: Loss

Given two disjoint datasets as D_L (labeled dataset),D_U (unlabeled dataset), the model as *f* is optimized by the conjunction of the supervised and unsupervised loss function defined as

$$\underset{\square}{min}\mathbb{E}_{(x,y)\sim P(X,Y)}[l(y, f(x))] +\ \beta\ \mathbb{E}_{(x,y_{ps})\sim P(X,Y_{ps})}[l_u(y_{ps}f(x))] \tag{3}$$

First term is estimated over a small labeled subset, and the unsupervised loss over the unlabeled subset. For this framework, the *lu* is defined to support the supervised objective on pseudo-labelling.

Methodology

Triplet Mixup Data Augmentation was proposed by [12], where they used this data augmentation technique to generate minor examples to fit tabular domains:

$$\tilde{x} =\ \lambda_i\, x_i + \lambda_j x_j + (1 - \lambda_i - \lambda_j)x_k \tag{4}$$

where they used all the data from the minority (vulnerable) class because their goal was to alleviate the data imbalance problem when using continuous data. In our case, the context is different since we do have continuous and categorical features. On the other hand, we propose Triplet Mixup to interpolate between data belonging from the same class to create positive samples in the hidden space, instead of doing so in the input space as [12]. As mentioned by [11], low probable samples may be produced by Mixup in the input space due to the multi-modality of the data and the categorical features. More specifically, given an encoder E, that contains f_T layers (T = total of layers), produces abstract representations of the input samples in the intermediate layers that are then interpolated with equation 4 to ensure high probable samples.

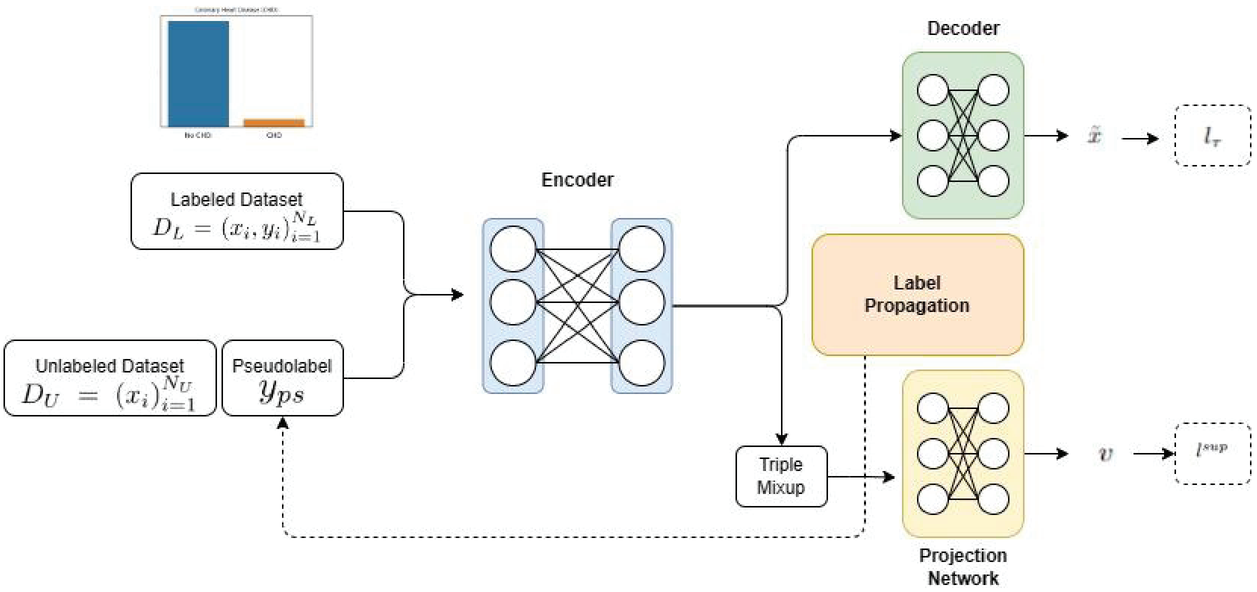


Figure 1. Contrastive Triple Mix Up in hidden space Framework.

Extending the definition of 1 for the interpolated intermediate layer samples generated by the encoder e that is composed of T layers *ft* ($t \in 1, ..., T$). The Triple Mix up create the interpolation in the hidden layers as:

$$h^t{}_{ijk} = \lambda_i h^t{}_i + \lambda_j h^t{}_j + (1 - \lambda_i - \lambda_j) h_k{}^t \tag{5}$$

Where λ is a scalar from a uniform distribution U(0, α) with $\alpha \in [0, 0.5]$.

From the equation (2) to maximize the distance for feature classes belonging to different classes and minimize the distance for same feature classes, the loss term is as described in [11].

Future Work.

The priority of this work is to study the influence of the proposed triple mix-up on the hidden space as an augmentation technique to attack the tabular imbalanced dataset, using the framework by the authors in [11]; given the actual contrastive learning loss used [20]; benefits from larger batch sizes and longer training. Also, we want to investigate the impact of applying data augmentation on medical dataset in an early stage [12] mixed with Triplet Mixup in the inner stages of the model. The main intention is to generate more data given a single X keeping a good generalization in the classification of the Coronary Heart disease problem.

Referencias

[1] M. Tan, R. Pang, and Q. V. Le, “EfficientDet: Scalable and efficient object detection,” in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2020, pp. 10 778–10 787.

[2] K. He, G. Gkioxari, P. Dollar, and R. Girshick, “Mask R-CNN,” ´ IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 2, pp. 386–397, 2020.

[3] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2020.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137–1149, 2017.

[5] Z. Q. Zhao, P. Zheng, S. T. Xu, and X. Wu, "Object Detection with Deep Learning: A Review," IEEE Transactions on Neural Networks and Learning Systems, vol. 30, no. 11, pp. 3212–3232, 2019.

[6] D. Snow, "DeltaPy: A Framework for Tabular Data Augmentation in Python," SSRN Electronic Journal, pp. 1–3, 2020.

[7] B. Sathianarayanan, Y. C. Singh Samant, P. S. Conjeepuram Guruprasad, V. B. Hariharan, and N. D. Manickam, "Feature-based augmentation and classification for tabular data," CAAI Transactions on Intelligence Technology, vol. 7, no. 3, pp. 481–491, 2022.

[8] G. Somepalli, M. Goldblum, A. Schwarzschild, C. B. Bruss, and T. Goldstein, "Saint: Improved neural networks for tabular data via row attention and contrastive pre-training," 6 2021. [Online]. Available: http://arxiv.org/abs/2106.01342

[9] J. Yoon, Y. Zhang, J. Jordon, and M. van der Schaar, "Vime: Extending the success of self-and semi-supervised learning to tabular domain," Advances in Neural Information Processing Systems, vol. 33, pp. 11 033–11 043, 2020.

[10] M. Hyun, J. Jeong, and N. Kwak, "Class-imbalanced semi-supervised learning," 2 2020. [Online]. Available: http://arxiv.org/abs/2002.06815

[11] S. Darabi, S. Fazeli, A. Pazoki, S. Sankararaman, and M. Sarrafzadeh, "Contrastive mixup: Self- and semi-supervised learning for tabular domain," 2021. [Online]. Available: http://arxiv.org/abs/2108.12296

[12] X. Li, L. Khan, M. Zamani, S. Wickramasuriya, K. W. Hamlen, and B. Thuraisingham, "Mcom: A semi-supervised method for imbalanced tabular security data" in IFIP Annual Conference on Data and Applications Security and Privacy. Springer, 2022, pp. 48–67.

[13] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," 10 2020. [Online]. Available: http://arxiv.org/abs/2011.00362

[14] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," 2 2020. [Online]. Available: http://arxiv.org/abs/2002.05709

[15] P. M. Tripathi, A. Kumar, R. Komaragiri, and M. Kumar, A Review on Computational Methods for Denoising and Detecting ECG Signals to Detect Cardiovascular Diseases. Springer Netherlands, 2022, vol. 29, no. 3. [Online]. Available: https://doi.org/10.1007/s11831-021-09642-2

[16] A. Subas, E. Alickovic, and J. Kevric, "Diagnosis of chronic kidney disease by using random forest," IFMBE Proceedings, vol. 62, no. 3, pp. 589–594, 2017.

[17] W. Deng, Z. Huang, J. Zhang, and J. Xu, "A Data Mining Based System for Transaction Fraud Detection," 2021 IEEE International Conference on Consumer Electronics and Computer Engineering, ICCECE 2021,pp. 542–545, 2021.

[18] D. Krishnani, A. Kumari, A. Dewangan, A. Singh, and N. S. Naik, "Prediction of coronary heart disease using supervised machine learning algorithms," IEEE Region 10 Annual International Conference, Proceedings/TENCON, vol. 2019-Octob, pp. 367–372, 2019.

[19] H. Yang, "Coronary heart disease historical data," 2022. [Online]. Available: https://dx.doi.org/10.21227/eapx-t883

[20] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," CoRR, vol. abs/2002.05709, 2020. [Online]. Available: https://arxiv.org/abs/2002.05709

[21] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," 2019. [Online]. Available: https://arxiv.org/abs/1911.05722

Tracking the trajectory of a swarm of mobile robots with a computer vision system

Seguimiento de la trayectoria de un enjambre de robots móviles con un sistema de visión por computadora

Andrés Jiménez-Mora[1], Kevin Morales-Paz[2], Juan Carlos Brenes-Torres[3], Rebeca Solís-Ortega[4], Cindy Calderón-Arce[5]

Jiménez-Mora, A; Morales-Paz, K; Brenes-Torres, J. C; Solís-Ortega, R; Calderón-Arce, C. Tracking the trajectory of a swarm of mobile robots with a computer vision system. *Tecnología en Marcha*. Vol. 37, special issue. August, 2024. IEEE International Conference on BioInspired Processing. Pag. 44-49.

https://doi.org/10.18845/tm.v37i7.7296

[1] Instituto Tecnológico de Costa Rica, Costa Rica.
andjm28@estudiantec.cr.
https://orcid.org/0009-0002-7446-832X

[2] Instituto Tecnológico de Costa Rica, Costa Rica.
kevinmoralespaz@gmail.com
https://orcid.org/0009-0006-6833-3479

[3] Escuela de Ingeniería en Mecatrónica, Instituto Tecnológico de Costa Rica, Costa Rica. Estudiante Doctorado en Automática, Robótica e Informática Industrial. Universidad Politécnica de Valencia, España.
juanbrenes@tec.ac.cr
https://orcid.org/0000-0001-6323-2173

[4] Escuela de Matemática, Instituto Tecnológico de Costa Rica, Costa Rica.
rsolis@tec.ac.cr
https://orcid.org/0000-0002-3065-8386

[5] Escuela de Matemática, Instituto Tecnológico de Costa Rica, Costa Rica.
ccalderon@tec.ac.cr
https://orcid.org/0000-0002-0077-225X

Keywords

Swarm robotics; computer-based vision systems; trajectory tracking; OpenCV

Abstract

Swarm robotics research uses a range of tools for evaluating the behaviors and metrics of robot collectives. One crucial tool involves the capability to track each robot's position and orientation at various intervals, enabling the reconstruction of individual robot poses and trajectories. Comprehensive analysis of swarm behavior hinges on the study of the collective trajectories of each robot within the group. This paper demonstrates the implementation of a computer vision system, utilizing a webcam and Python scripts, to effectively track a mobile robot group within a swarm. This shows the feasibility of developing such research tools using commonplace computing equipment. The design and development of the vision system, including a detailed calibration procedure, robot identification methods, and practical examples, are also shown. Furthermore, it offers an exhaustive explanation of the robot tracking process. Experimental trials with three robots validate the system's ability to extract images from video feeds and accurately identify each robot. Subsequently, after image processing, the system generates a dataset encompassing image numbers, robot IDs, x and y positions, and orientations.

Palabras clave

Robótica de enjambres; sistemas de visión por computadora; seguimiento de trayectorias; OpenCV.

Resumen

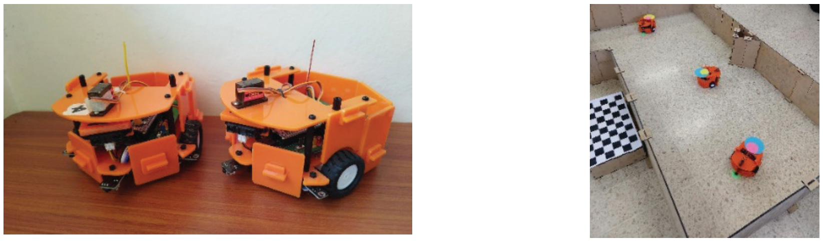
En el campo de la robótica de enjambres se utilizan una variedad de herramientas para evaluar los comportamientos y las métricas de los colectivos de robots. Una herramienta crucial implica la capacidad de rastrear la posición y orientación de cada robot en varios intervalos, lo que permite la reconstrucción de posturas y trayectorias seguidas por los mismos. El análisis exhaustivo del comportamiento de los enjambres depende del estudio de las trayectorias colectivas de cada robot dentro del grupo. Este artículo demuestra la implementación de un sistema de visión por computadora, que utiliza una cámara web y scripts de Python, para rastrear de manera efectiva un grupo de robots móviles dentro de un enjambre. Esto muestra la viabilidad de desarrollar tales herramientas de investigación utilizando equipos informáticos comunes. Adicionalmente, se muestra el diseño y desarrollo del sistema de visión, incluido un procedimiento de calibración detallado, métodos de identificación de robots y ejemplos prácticos. Además, ofrece una explicación exhaustiva del proceso de seguimiento del robot. Las pruebas experimentales con uno y tres robots validan la capacidad del sistema para extraer imágenes de videos e identificar con precisión cada robot. Posteriormente, después del procesamiento de imágenes, el sistema genera un conjunto de datos que abarca números de imágenes, ID de robots, posiciones (x, y) y orientaciones.

Introduction

In the subject of robotics, swarm robotics draws inspiration from the self-organizing systems found in nature, such as social insects, schools of fish, and flocks of birds. These systems exhibit collective behavior arising from simple local interactions [1]. This paper introduces the

PROE project [2], supported by the Costa Rica Institute of Technology, which focuses on the implementation of a swarm robotics prototype for exploring static scenarios and optimizing routes [3, 4, 5].

The project developed a group of robots called Atta-Bots. They are equipped with sensors to identify obstacles and transmit their position data (Figure 1a). The robots are capable of transmitting data about their position and orientation, which is captured by the data processing system to create a map of the area covered by the swarm. Using this map, optimal routes can be defined within this environment. In the case study, the environment is controlled and consists of a regular surface of granite mosaic, upon which a scenario with reconfigurable MDF walls is mounted (Figure 1b).



(a) Atta-Bots

(b) Experimental configuration

Figure 1. Project description.

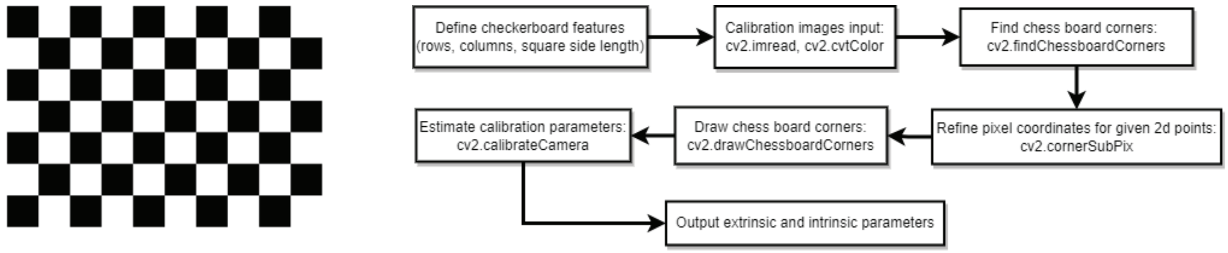
Materials and methods

Software tools

The foundation of the current vision system is the programming language Python and its library for Open Computer Vision (OpenCV), which consists of a collection of code intended for real time artificial vision, containing several functions for image interpretation. The library OpenCV is open source, it was first developed by Intel Corporation, and it is currently on its release version 4.8.0 [6].

Calibration

The purpose of this process is to determine the intrinsic and extrinsic parameters of a camera to accurately translate a three-dimensional point into a two-dimensional projection within the camera coordinate system. In this case, the plane pattern method was applied to perform the camera calibration [7, 8]. The checkerboard shown in figure 2a was used as the reference pattern by placing it in several spots within the experimental scenario and consequently, saving the images for further processing. The flowchart from figure 2b shows in a high level the implementation of the calibration process using OpenCV functions.


Define checkerboard features (rows, columns, square side length)
Calibration images input: cv2.imread, cv2.cvtColor
Find chess board corners: cv2.findChessboardCorners
Refine pixel coordinates for given 2d points: cv2.cornerSubPix
Draw chess board corners: cv2.drawChessboardCorners
Estimate calibration parameters: cv2.calibrateCamera
Output extrinsic and intrinsic parameters

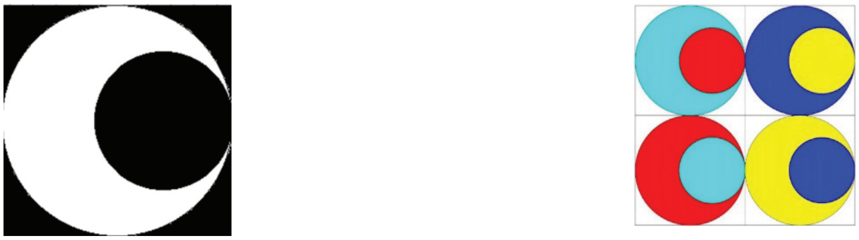
(a) Calibration pattern

(b) Calibration process

Figure 2. Camera calibration.

Identifier design and detection

The developed implementation tracks the movement of the robots by identifying the displacement between pairs of images taken at the initial moment and the subsequent moment. To locate each robot in every image, an identifier composed of 2 circles, one large and one small, was used (Figure 4a). The designed identifier has a side length of 10 cm, which is the maximum allowed by the robot's dimensions. The size of the two circles was maximized by fitting the smaller circle inside the larger one, and different colors were used for each circle to differentiate them (Figure 4b).



(a) Identifier base design

(b) Color identifier

Figure 4. Identifier pattern designed for the application

The chosen colors must stand out from the rest of the objects in the image. Therefore, striking and easily identifiable colors were selected, including blue, cyan, dark green, red, magenta, yellow, and orange. Using these colors, pairs were created, and the modified identifier designed earlier was used. The experimental setup with these identifiers is depicted in Figure 5.

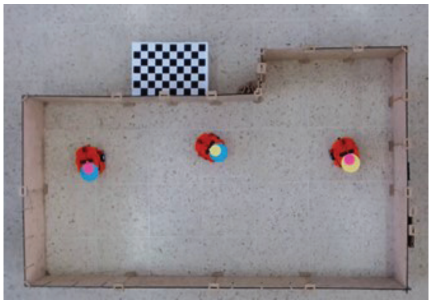


Figure 5. Camera view of the robots with color identifiers.

For the detection of the circles that make up the identifier, the Circular Hough Transform is used, with the utilization of minimum and maximum radius parameters for searching in the image, expressed in pixels.

Robot tracking process

The robot tracking process involves the initial capture of robot images, followed by a comprehensive analysis of these visual data. Through image processing function from OpenCV, the system extracts valuable information from the images, enabling the identification and localization of robots. The process steps are illustrated in the flowchart from figure 6, provide a brief explanation of each step within the robot tracking process. At the end of the processing of all the specified images, the program generates a text file (txt) where it reports the displacement between frames for each robot, and the angle that the robot has with respect to the horizontal axis.

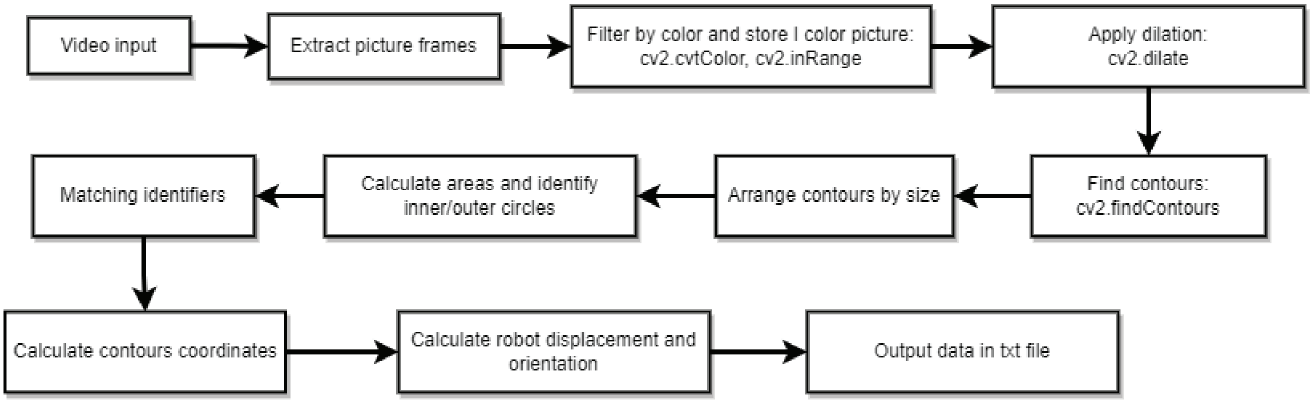
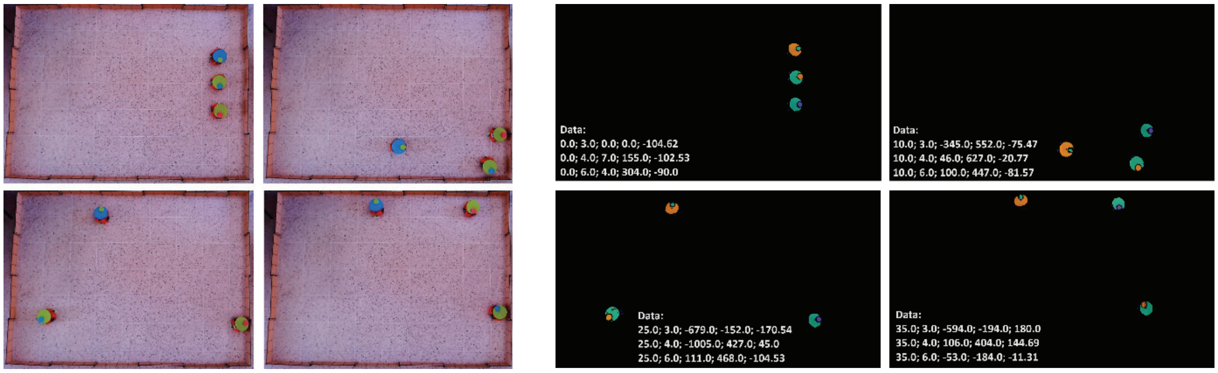

Video input
Extract picture frames
Filter by color and store I color picture: cv2.cvtColor, cv2.inRange
Apply dilation: cv2.dilate
Find contours: cv2.findContours
Arrange contours by size
Calculate areas and identify inner/outer circles
Matching identifiers
Calculate contours coordinates
Calculate robot displacement and orientation
Output data in txt file

Figure 6. Diagram of the robot tracking process.

Results

The validation methodology was oriented to check if the system is able to detect and calculate the trajectory of the robot. Consequently, robots were placed on a controlled and closed scenario of an approximate dimension of 2x3 meters. Scenario walls were adaptable panels fabricated from medium-density fiberboard (MDF). A digital camera (webcam with a 1080p resolution) was placed on the ceiling, looking downwards. This environment presents abundant artificial light and smooth floor surface with a light granite color. The experiment was done using three robots. The swarm explored the scenario by employing a random walk behavior for about a minute. The present challenge was not only to track a robot and its trajectory, but to distinguish between robots. A video was recorded and a total of 48 pictures were extracted from it (see Figure 8a). Those pictures were successfully processed by employing the computer vision system. Each robot was distinguished and identified correctly by the system, and its trajectory tracked successfully. A sample of the processed pictures and its output data is shown on figure 8b.



(a) Sample of pictures

(b) Processed pictures

Figure 8. Experiment with three robot. Data shown represents: picture number, robot ID, position on x axis (mm), position on y axis (mm), and robot orientation (degrees). Data is over-imposed on picture for explanatory purposes.

Conclusions and future work

The behavior of a robot swarm can be comprehensively analyzed by studying the collective trajectories of each individual robot within the group. This paper has shown the implementation of a computer vision system, using a webcam and a set of Python scripts, to effectively track a group of mobile robots within a robotic swarm, using common computing equipment. Nevertheless, experiments conducted with three robots demonstrated the system's successful extraction of images from a video feed, accurately identifying each robot. Subsequently, after processing all the images, the system generated a dataset comprising image numbers, robot IDs, x and y positions, as well as orientation. Future research endeavors could investigate the application of this tool with a live video feed and real-time calculation of robot trajectories.

Acknowledgements

The authors would like to thank the Research and Outreach Department of the Costa Rica Institute of Technology (VIE, ITCR), for their support of the PROE project (code VIE 5402-1440-5101). These results are part of that project.

References

[1] M. Dorigo, G. Theraulaz, and V. Trianni, "Swarm robotics: Past, present, and future [point of view]," *Proceedings of the IEEE*, vol. 109, no. 7, pp. 1152–1165, 2021.

[2] J. C. Brenes-Torres, "jcbrenes/proe: Proyecto proe: Implementación de un prototipo de enjambre de robots para la digitalización de escenarios estáticos y planificación de rutas óptimas."

https://github.com/jcbrenes/PROE, 2021

[3] C. Calderón-Arce and R. Solís-Ortega, "Swarm robotics and rapidly exploring random graph algorithms applied to environment exploration and path planning," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 5, 2019.

[4] R. Solis-Ortega and C. Calderon-Arce, "Multiobjective problem to find paths through swarm robotics," in *Proceedings of the 2019 3rd International Conference on Automation*, *Control and Robots*, 2019, pp. 12–21.

[5] J. C. Brenes-Torres, F. Blanes, and J. Simo, "Magnetic trails: A novel artificial pheromone for swarm robotics in outdoor environments," *Computation*, vol. 10, no. 6, p. 98, 2022.

[6] Intel, Open-Source Computer Vision Library. Reference Manual, 2001.

[7] W. Qi, F. Li, and L. Zhenzhong, "Review on camera calibration," in *2010 Chinese Control and Decision Conference*, 2010, pp. 3354–3358.

[8] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330–1334, 2000.

Exploration and selection of LLM models for financial text simplification

Exploración y selección de modelos LLM para la simplificación de texto financiero

Bertha C. Brenes-Brenes[1], Saul Calderón-Ramírez[2]

Brenes Brenes, B. C; Calderón-Ramírez, S. Exploration and selection of LLM models for financial text simplification. *Tecnología en Marcha*. Vol. 37, special issue. August, 2024. IEEE International Conference on BioInspired Processing. Pag. 50-56.

https://doi.org/10.18845/tm.v37i7.7297

[1] Student. Instituto Tecnológico de Costa Rica, Costa Rica.
berthabb@estudiantec.cr
https://orcid.org/0009-0008-1303-7263

[2] Computer Scientist, Instituto Tecnológico de Costa Rica, Costa Rica.
sacalderon@itcr.ac.cr
https://orcid.org/0000-0001-9993-4388

Keywords

LLM models; simplification; SARI; BLEU; AI; Llama; financial data.

Abstract

This research is dedicated to the simplification of Spanish-language financial texts to enhance accessibility for screen readers. We present a qualitative and quantitative analysis of the text simplification process, employing a set of Spanish simplification rules and metrics. Our study evaluates the outcomes resulting from the application of three distinct financial datasets to four pre-trained models. The primary objective is to identify the most effective models for text simplification and determine those warranting further investment through fine-tuning and training. This study contributes to improving the accessibility and comprehensibility of financial documents for individuals with visual impairments.

Palabras clave

Modelos LLM; simplificación; SARI; BLEU; IA; datos financieros.

Resumen

Esta investigación está dedicada a la simplificación de textos financieros en español para mejorar la accesibilidad de los lectores de pantalla. Presentamos un análisis cualitativo y cuantitativo del proceso de simplificación de textos, empleando un conjunto de reglas y métricas de simplificación en español. Nuestro estudio evalúa los resultados obtenidos de la aplicación de tres conjuntos de datos financieros a cuatro modelos pre entrenados. El objetivo principal es identificar los modelos más eficaces para la simplificación de textos y determinar aquellos que justifican una mayor inversión para el fine-tunning y el entrenamiento. Este estudio contribuye a mejorar la accesibilidad y comprensibilidad de los documentos financieros para las personas con discapacidad visual.

Introduction

Large Language Models (LLMs) are powerful machine learning models that leverage extensive training data to understand, translate, generate,simplify or summarize text. These models offer versatility in their application and training methods. However, their usage typically demands significant computational resources due to their memory and processing requirements, often surpassing the capabilities of a single computer.

One prominent type of an LLM Modelos is the Transformer [1]. Transformer architecture excels at analyzing text to comprehend its context and generate human-like responses. It operates by breaking down input text into tokens and representing them as vectors, capturing both their meaning and context. This allows the users to interact with the model using natural language queries and instructions, harnessing the context for generating meaningful responses [2].

The central to the effectiveness of LLMs is the quality and relevance of the training dataset. In this study, we focus on a dataset consisting of financial documents, specifically tailored for individuals with visual impairments, already develop by another research team

Our primary objective is to simplify Spanish-language financial text to enhance the accessibility of paperwork in screen readers for visually impaired people. This is a priority in our project, because we stand with the equal accessibility of all the people. We can define simplification as the replacement of complex words to simpler words to make the sentences more simple and easy to understand. In this process it is needed to delete, add or replace some words. One short example of simplification is:

Complex text: *"In recent years, specifically in recent decades, communities have become a fundamental pillar of their own economy, of countries and even of the international economy."*

Simpler text: *"In recent decades, communities have become a pillar of the local and international economy."*

This paper aims to present a qualitative and quantitative analysis using different types of spanish simplification rules, and utilizing a range of simplification metrics to evaluate the results obtained from three distinct financial datasets and four pre-trained models. The goal is to identify the most effective models for simplifying text and determine which ones are worthy of further investment in terms of fine-tuning and training.

Theoretical framework

For this project we focus on the primary exploration of the models and their performance using metrics. Is important to explain the concepts of the models that we are using:

Llama Model: LLM model developed by Meta, we are using Llama 2-13bf. which is optimized for dialogue use cases. **Pegasus Model**: LLM Model developed by Google, is pre-training with Extracted Gap-sentences for Abstractive SUmmarization Sequence-to-sequence models.

FastChat Model: LLM Model developed by LMSYS, is an open platform for training, serving, and evaluating large language model based chatbots. **Alpaca Model**: LLM Model developed by Stanford University, a model fine-tuned from the LLaMA 7B model on 52K instruction-following demonstrations.

Mectrics

After all the text has been applied to the models mentioned above it is necessary to evaluate the simplification. It was necessary to use metrics that work efficiently in Spanish and for a better understanding of their objective in the simplification the metrics will be presented divided on the category they work to evaluate the quality of each simplification.

First we have the metrics that evaluate the quality of text simplification:

- **BLEU**:This metric does not have a specific equation because it will depend on the n-gram which is actually a widely used concept from regular text processing and is not specific to NLP or Bleu Score. '"The BLEU metric is always a number between 0 and 1'[5]. This result value indicates how similar the candidate text is to the reference texts, with values closer to 1 representing more similar texts.
- **SARI**: The metric compares the predicted simplified sentences against the reference and the source sentences. Is defined by .The range of values goes from 0 to 100, the higher the value, the better the simplification is.

The second category is the lexical simplicity metrics for natural language. All these metrics work for Spanish text.

- **Fernandez huerta**: This metric is defined by , where μ is the average number of syllables per word and Fμ the average number of words per phrase. It goes from 0 to 100, where 100 is easier to read.

- **Szigriszt pazos**: This metric is defined by . It goes from 0 to 100 where 0 is the hardest. And contrary as in the Fernandez huerta, it explains how hard it is to read a sentence.
- **Gutierrez poloni**: proposed as a novel readability formula from scratch for Spanish, where: L is the total number of characters. A higher value indicates a higher readability.
- **Crawford**: Is defined by the formula , OP, the number of sentences per hundred words; SP, the number of syllables per hundred words, This formula returns an estimate of the years of schooling required to understand the text[3] The lower the value indicates an easy readability.

The third category evaluates simple to complex text comparison. In addition to the following metrics, various features like syllables, word count, polarity (polo), monosyllable count, exact copies, and additions/deletions proportion are used to evaluate.

- Compression ratio: Calculates the ratio of the number of characters between the simple and complex sentences, is defined as . The lower this metric, the more simple it is.
- Levenstein similarity: this metric measures the Levenshtein distance between two text segments as the minimum number of single character edits (insertions, deletions or substitutions). A lower value indicates more lexical transformations.
- Sentence split: Is defined as .It corresponds to the ratio between the number of sentences in the simple text segment and the number of sentences in the complex text segment , thus: The lower this metric, the more complex the sentences is
- Lexical complexity score: Is a task in NLP and computational linguistics that involves assessing the level of complexity of the vocabulary and word usage in a given text[4]. A higher value indicates a good simplification and is easier to understand.

Methodology

For this research project, we will use a dataset that has been developed for a research group. They extract information from 4 financial books and divide this text in segments, then evaluate manually the complexity of each segment in order to create a several dataset where the biggest count is more than 5000 complex texts, this dataset also contain a manual simplification and a gpt3 simplification.We can understand this better with the following example:

Book	Complex	Simple
15654_LibroBAC.pdf	Este se da cuando el asegurado llega donde el médico, recibe la atención, efectúa el pago directamente y después solicita a la aseguradora el reembolso.	Esto pasa cuando el asegurado recibe la atención médica, efectúa el pago y después solicita a la aseguradora el reembolso.

Therefore we structured our approach to this dataset into three distinct stages:

Stage 1: Model Exploration: In this initial phase, we began with an exploration of various language models. In the research group, there was an initial exploration with 13 complex texts using models such as GPT-3, Llama13, Alpaca 13b, FastChat, and mt5 in the web chat.lmsys. Our exploration aimed to delve deeper into these models by accessing their Hugging Face pre-trained versions in Spanish so we can explore their usage in python, how it responds to the prompt instruction and the computer resource usage.

a. **Model Selection:** We identified one or more models for each of the four mentioned. Our selection criteria included evaluating the quality of documentation, scientific or blog references, Python compatibility, and computational resource requirements.

b. **Text Testing:** We used 13 complex texts to assess the four selected models. This involved applying the BLEU and SARI metrics to generate a range of scores for each of the 13 texts.

c. **Manual Validation:** Additionally, we manually validated the simplifications based on established guidelines. Each text was rated on a scale from 1 to 5, with 5 representing the highest quality simplification.

Stage 2: Large-Scale Testing: In the second stage, we scaled up our analysis by applying a dataset of over 5,000 complex texts to each of the models. The following steps were undertaken:

a. **Metrics Application:** We applied all the metrics mentioned in the theoretical framework, including SARI and BLEU, to assess the quality of simplifications.

b. **Replication:** To ensure robust results, we conducted more than five replications for each text and model, allowing for a more detailed and specific evaluation.

Stage 3: Shorter Dataset Analysis: The final stage involved the application of a shorter dataset containing fewer than 3,000 complex texts. This dataset was curated for its quality and relevance but kept the same structure that was already defined

Results

For the first stage The models Alpaca, Llama, and FastChat demonstrated positive performance in simplifying Spanish text, and each had their moments to stand out in different contexts. The application of rules was more consistent and accurate compared to the Pegasus model. In table 1 we can see the values apply to the simplification and it shows how the FastChat models represent the best. I consider that pegasus could improve if we use a specialized dataset and better training, because its main failure today is the number of words with grammatical and incoherent errors, so perhaps it will improve in a next stage with fine-tuning.

Table 1. Results from the manual evaluation. The values range from 1 to 5, the higher is 5.

Model	Total
FastChat	4.75
Llama	4.08
Alpaca	3.75
Pegasus	2.75

For the second and third stage we have the average of the results for BLEU and SARI. According with the characteristics of the implementation of SARI and BLEU we verify the metrics with different references, as we can see on the following tables

Table 2. quality of text simplification metrics in SARI and BLEU respectively.

Num. refs	Alpaca	Llama	FastChat	Pegasus
2.0	21.40	21.50	24.73	30.35

Num. refs	Alpaca	Llama	FastChat	Pegasus
3.0	21.33	21.42	24.73	30.57
4.0	21.06	21.16	24.79	31.34
5.0	21.19	21.31	25.76	31.75

Num. refs	Alpaca	Llama	FastChat	Pegasus
2.0	55.54	59.32	71.47	18.07
3.0	56.18	59.60	71.47	16.73
4.0	55.54	58.79	77.63	16.76
5.0	56.01	59.25	78.18	16.56

Table 3. lexical simplicity metrics, the higher the values the better, except for the Szigriszt_pazos.

Metrics	Alpaca	Llama	FastChat	Pegasus
Fernandez huerta	77.13	79.15	76.22	99.12
Szigriszt pazos	73.95	75.94	72.78	95.80
Gutierrez poloni	35.77	37.51	36.47	46.37
Crawford	3.86	3.71	3.89	2.30

Table 4. simple to complex text comparison metrics. the higher the values the better, except for the compression ratio and Levenstein similarity.syllable, word and polly are just informative, addition and deletion.

Metrics	Alpaca	Llama	FastChat	Pegasus
Compression ratio	1.09	1.01	1.00	0.43
Levenstein similarity	0.96	1.00	1.00	0.55
Sentence split	1.00	1.31	1.00	1.03
Lexical complexity	9.65	9.65	9.73	9.41
Additions proportion	0.08	0.02	0.00	0.07
Deletions proportion	0.00	0.00	0.00	0.63
syllable count	48.12	46.21	50.11	15.17
word count	28.59	27.68	29.53	9.41
poly count	5.11	5.11	5.91	1.54
SBert	0.96	1.0	1.0	0.72

Discussion

The comprehensive analysis of our research findings, as presented in Table 2, underscores the strong performance of the FastChat model in simplifying Spanish text. Pegasus, while showing promise in SARI, exhibited a substantial deficit in BLEU scores. Nevertheless, it excelled in lexical simplification, which is an important finding. In Table 4, where we compared the transformation of complex to simple text, FastChat once again demonstrated its prowess, even without additions or deletions in the simplification process. Alpaca and Llama also delivered commendable results, positioning them not far behind FastChat. Taking all these values into consideration, we can conclude that FastChat has demonstrated a better performance, making it a prime candidate for the fine-tuning process. In contrast, Pegasus may warrant reconsideration. It's noteworthy that these results were obtained within the context of an exploration approach without fine-tuning. This offers an optimistic outlook for the subsequent stages of our research, with the potential for even more impressive outcomes.

References

[1] P. Menon. (2023). Introduction to Large Language Models and the Transformer Architecture

[2] V.Chaudhary.(2023).Transformers and LLMs: The Next Frontier in AI, ur(https://www.linkedin.com/pulse/transformers-llms-next-frontier-ai-vijay-chaudhary/)

[3] Legible. Fórmula de Crawford, url(https://legible.es/blog/formula-de-crawford/)

[4] K.North,M.Zampieri, M.Shardlow.(2023). Lexical Complexity Prediction: An Overview, url(https://dl.acm.org/doi/10.1145/3557885)

[5] K. Doshi(2021). Foundations of NLP Explained — Bleu Score and WER Metrics, https://towardsdatascience.com/foundations-of-nlp-explained-bleu-score-and-wer-metrics-1a5ba06d812b *Repo: https://github.com/BerthaBrenes/Text-Simplification-with-LLM*

Antioxidant and antibacterial extracts from rambutan (*Nephelium lappaceum*) skins: Exploring the Potential of Transforming Agricultural Byproducts into Functional Supplements

Extractos antioxidantes y antibacteriales de las cáscaras del rambután (*Nephelium lappaceum*): explorando el potencial de transformación de subproductos agroindustriales en suplementos funcionales

Victor Álvarez-Valverde[1], Carlos Alfaro-Zúñiga[2], Andreia Passos-Pequeno[3], Yendri Carvajal-Miranda[4], Jorengeth Abad Rodríguez-Rodríguez[5], Gerardo Rodríguez[6], Pablo Jiménez-Bonilla[7]

Álvarez-Valverde, V; Alfaro-Zúñiga, C; Passos-Pequeno, A; Carvajal-Miranda, Y; Rodríguez-Rodríguez, J. A; Rodríguez, G; Jiménez-Bonilla, P. Antioxidant and antibacterial extracts from rambutan (nephelium lappaceum) skins: exploring the potential of transforming agricultural byproducts into functional supplements. *Tecnología en Marcha*. Vol. 37, special issue. August, 2024. IEEE International Conference on BioInspired Processing. Pag. 57-67.

https://doi.org/10.18845/tm.v37i7.7298

[1] Laboratorio de Fitoquímica (LAFIT), Escuela de Química, Universidad Nacional, Heredia, Costa Rica.
victor.alvarez.valverde@una.ac.cr
https://orcid.org/0000-0001-6007-9150

[2] Ministerio de Agricultura y Ganadería, MAG. Costa Rica.
carlos.edo.28@gmail.com

[3] Escuela de Medicina Veterinaria, Universidad Nacional, Heredia, Costa Rica.
andreia.passos.pequeno@una.ac.cr

[4] Laboratorio de Fitoquímica (LAFIT), Escuela de Química, Universidad Nacional, Heredia, Costa Rica.
yendry.carvajal.miranda@una.ac.cr
https://orcid.org/0000-0002-4460-3466

[5] Escuela de Ciencias Biológicas, Laboratorio de Biotecnología Microbiana, Universidad Nacional, Heredia, Costa Rica.
jorengeth.rodriguez.rodriguez@una.ac.cr
https://orcid.org/0000-0001-8452-8256

[6] Laboratorio de Fitoquímica (LAFIT), Escuela de Química, Universidad Nacional, Heredia, Costa Rica.
garodriguezr57@gmail.com
https://orcid.org/0000-0002-1224-0277

[7] Laboratorio de Fitoquímica (LAFIT), Escuela de Química, Universidad Nacional, Heredia, Costa Rica.
pablo.jimenez.bonilla@una.ac.cr
https://orcid.org/0000-0002-5786-9845

Keywords

Rambutan; antioxidants; antibacterial; *Nephelium lappaceum*; functional supplements.

Abstract

Rambutan (*Nephelium lappaceum* L.) is a tropical fruit characterized by its oval shape and hairy skin, primarily valued for its juicy pulp. The peels, constituting 45% of the fruit's weight, are a source of valuable compounds like geraniin, ellagic acid, and quercetin. These peels possess antimicrobial properties effective against various bacteria, making them suitable for food preservation and packaging. Additionally, rambutan extracts hold promise as supplements in animal feed, enhancing growth and reducing methane production. This research delves into the antioxidant and antimicrobial attributes of diverse rambutan varieties. The skin (exocarp) of rambutan fruits from three Costa Rican cultivars -Creole, Rongrein, and Yellow- were collected and processed. Total polyphenolic content (TPC), proanthocyanidins, antioxidant activity, geraniin content, and antimicrobial activity were determined for the three varieties. Also, proanthocyanidin-enriched fractions from rambutan extracts were generated and analyzed. The results revealed TPC and antioxidant activity variations among different rambutan varieties and harvest years. All rambutan extracts displayed antimicrobial activity. In conclusion, the research underscores the rich antioxidant content in rambutan peels, irrespective of the variety, and underscores their potential for use in both human and animal nutrition due to their chemical composition.

Palabras clave

Rambután; antioxidantes; antibacterial; *Nephelium lappaceum*; suplementos funcionales.

Resumen

El rambután (*Nephelium lappaceum L.*) es una fruta tropical caracterizada por su forma ovalada y su cáscara peluda, de la cual principalmente se aprovecha su pulpa jugosa. Las cáscaras, que constituyen el 45% del peso de la fruta, y son una fuente de compuestos valiosos como la geranina, el ácido elágico y la quercetina. Estas cáscaras poseen propiedades antimicrobianas efectivas contra varias bacterias, lo que las hace adecuadas para la conservación y el envasado de alimentos. Además, los extractos de rambután prometen ser suplementos en la alimentación animal, mejorando el crecimiento y reduciendo la producción de metano. Esta investigación se adentra en los atributos antioxidantes y antimicrobianos de diversas variedades de rambután. Para este efecto, se recolectaron y procesaron las cáscaras (exocarpo) de frutas de rambután de tres variedades costarricenses: Criolla, Rongrein y Amarillo. Se determinó el contenido total de polifenoles (TPC), proantocianidinas, actividad antioxidante, contenido de geranina y actividad antimicrobiana de las tres variedades. Además, se generaron y analizaron fracciones enriquecidas de proantocianidinas a partir de extractos de rambután. Los resultados revelaron variaciones en el TPC y la actividad antioxidante entre las diferentes variedades de rambután y los años de cosecha. Todos los extractos de rambután mostraron actividad antimicrobiana. En conclusión, la investigación resalta el rico contenido antioxidante en las cáscaras de rambután, independientemente de la variedad, y subraya su potencial uso tanto en la nutrición humana como animal debido a su composición química.

Introduction

The rambutan, scientifically known as *Nephelium lappaceum* L., is a tropical fruit classified within the Sapindaceae family [1]. Rambutan is a hairy oval-shaped fruit. It is composed of peel, pulp, seed, and embryo. The peels of this fruit are green, yellow, or red. The pulp is used for human consumption, and other parts are generally considered waste. Rambutan peels account for 45% of the fruit weight [2], and the peels are rich in geraniin, corilagin, rutin, ellagic acid, quercetin, and total phenolic compounds (TPC) [3]. Geraniin is an ellagitannin used as a food additive. Several health benefits have been associated with geraniin consumption [4]. It is a strong antioxidant compound, and it is reported to have antihypertensive, antiviral, antidiabetic, antihyperglycemic, and hepatoprotective activities [5].

Chemical composition varies depending on the variety utilized [6], plant maturity [7], and climate. Rambutan is a fruit from Asia, and it is produced in several tropical countries all over the world. The production of this fruit in Costa Rica is concentrated in the southern region. The commercialized rambutan fruit is a mixture of the available varieties.

Rambutan peel extracts are antimicrobials. Phuong *et al* [3] found antimicrobial activity against *Salmonella enteritidis, Pseudomonas aeruginosa, Escherichia coli, Staphylococcus aureus, Lactobacillus plantarum, Vibrio campbellii,* and *Listeria monocytogenes.* Interestingly, the same authors found no difference between the antimicrobial activity against antibiotic-resistant and non-resistant strains of *S. enteritidis* when rambutan peel extract of 100 µg GAE/mL concentration or higher is evaluated. Polyphenolic compounds are reported to be active against antibiotic-resistant bacteria [8] because their mechanism of action is different. The antimicrobial properties of rambutan extracts have been utilized for food preservation and food packing [9, 10]. *N. lappaceum* extracts have also been studied as an animal feed supplement, with no negative impact on animal microbiota composition [11]. Nonetheless, it reduces the methane production in beef and promotes better growth performance and immune response in Nile Tilapia [12], and catfish [13]. In this study, a preliminary assessment of the main antioxidant compounds and their antioxidant and antimicrobial activity was done. Our aim is to study the properties of different varieties from two different harvest times, as potential supplements for human or animal feeding applications.

Materials and methods

Biological materials

We collected the exocarp of the fruit from three rambutan cultivars (*Nephelium lappaceum* L) produced in Costa Rica. Creole, Rongrein, and Yellow varieties were sampled. Fruit was collected from the Pérez Zeledón canton in the province of San Jose, located at latitude 9°18’31.84” N, longitude 83°40’14.54” W. Fruits were collected during 2013 (samples R1), 2014 (R2), and 2015 (R3). The sampling region was selected because most rambutan producers are located there. The exocarp (skin) was separated from the pulp and the seed. Samples were stored in coolers after collection, immediately transported to the lab to be frozen, and then dehydrated in a Freeze-dryer 2.5 L plus (from Labconco Corp, MO, USA). The dried samples were ground in a medium-sized Wiley blade mill to a 1 mm (from Thomas Scientific, NJ, USA).

Optimization of the extraction procedure

The solvent type and number of extraction cycles were optimized. The solvents to be tested include acetone:methanol:5% HCl (4:4:2), acetone:5% HCl (7:3), acetone:ethanol:5% HCl (4:5:1), and 5% HCl in 95% ethanol. A composite sample comprising all cultivars in the same proportion was used. 75 mg of dry and ground sample was extracted with 3x3mL of the solvent.

In each extraction, the sample was sonicated with the solvent for 5 minutes. Then. the tubes were centrifuged at 400 g for 5 minutes, and the supernatant liquid was decanted. The three extractions were combined and adjusted to 10 mL final volume and analyzed using Folin-Ciocalteau´s protocol, as described below.

Then, optimal extraction cycles were optimized with the most efficient solvent. 5 test tubes, containing 75 mg of the sample were extracted using 1x2mL, 2x2mL, 3x2mL, 4x2mL, and 5x2mL, respectively. All tubes were adjusted to 10 mL and analyzed using Folin-Ciocalteau´s protocol.

Determination of Total Polyphenolic Compounds (TPC)

Peel material was directly extracted with the optimized extraction procedure. 75 mg of previously lyophilized and ground samples were extracted using three 3 mL aliquots of 95% ethanol acidified to 5%. The procedure was repeated 3 times. Subsequently, the combined 9 mL obtained was mixed to a final volume of 10 mL in an appropriate flask using acidified ethanol.

Three replicates of this procedure were prepared. Finally, the assessment of the TPC was conducted using the Folin-Ciocalteu colorimetric method, as previously described in our report [14]. Briefly, 30 µL of each of the previously prepared samples was mixed with 200 µL of water in a 96-well microplate for analysis. Subsequently, 15 µL of Folin-Ciocalteu reagent and 50 µL of a 20% sodium carbonate solution were added to the microplate. The plate was then incubated for 20 minutes with agitation at 40 °C in a Synergy HT Multi-Detection Microplate Reader (BioTek Instruments). After the incubation, the absorbance was measured at 755 nm, against standard solutions of 0.000, 0.020, 0.040, 0.060, 0.080, and 0.120 mg/mL gallic acid.

Rambutan proanthocyanin standards (rPAC) were also analyzed using Folin-Ciocalteau´s method. 1 mg of rPAC powdered was dissolved into 1 mL of ethanol and analyzed as described before.

Determination of antioxidant activity by the DPPH method

The antioxidant activity of DPPH (2,2 Diphenyl-1-picrylhydrazyl) was determined following the method described by Bondet [15], in a 96-well microplate. 30 µL of methanol:water (80:20) was used as blank. Wells designated as standards (STD) utilize 30 µL of a 0.0215 mg/mL gallic acid standard, and the wells designated for the samples (SPL) were filled with 30 µL of the sample, instead. Subsequently, 270 µL of the 0.042 mg/mL DPPH solution was added to each well, and the plate was placed in the microplate reader. After 30 minutes, the absorbance was measured at a wavelength of 515 nm.

Determination of antioxidant activity by ORAC method

Oxygen radical absorbance capacity was determined using the method described by Brescia [16]. 150 µL of 4x10^{-6} mM sodium fluorescein solution was added to all the wells. 25 µL of the samples were added to the wells. Then, the plate was incubated for 20 minutes at 37 °C in the microplate reader. After this incubation period, the plate was removed from the reader, and 25 µL of 2,2'-azobis-2-methyl-propanimidamide dihydrochloride (AAPH) solution 153 mM was immediately added to all the experimental wells. The microplate was placed back into the reader, and readings were taken every minute for a total of 60 readings per well. Fluorescence was measured at 485 nm (excitation) and 528 nm (emission), against 0, 25, 50, 75, and 100 µM Trolox (6-hydroxy-2,5,7,8-tetramethylchroman-2-carboxylic acid) standard curve. 75 mM phosphate buffer (pH 7.4) was used as a solvent for all solutions and blank.

Preparation of proanthocyanidin-enriched fractions of rambutan (rPAC)

Seven grams of dried rambutan sample were weighed, and five extractions of 50 mL each were performed using 95% ethanol. The obtained extracts were filtered and concentrated using a rotary evaporator under reduced pressure and 40°C. 5 mL ethanol 95% was added and centrifuged at 0°C and 11,000 rpm for 10 minutes. After centrifugation, the supernatant fluid was added to a column containing Sephadex LH-20, and eluted by adding ethanol: methanol in a 1:1 (v/v), and then acetone:water in a 7:3. Three fractions were collected, concentrated, freeze-dried, and analyzed for proanthocyanidins (PAC) (method described below). The fraction with the highest PAC concentration is considered the rPAC. They are stored at -18°C until use.

Determination of Proanthocyanidins

Three replicates and three repetitions each were done. 70 µL of sample and 210 µL of 0.1% (w/v) DMAC solution were mixed into each well. Immediately, the plate was placed into the reader and stirred for 10 seconds at 600 revolutions per minute (rpm). Then, the absorbance was measured at 640 nm every 30 seconds for one hour. Blank solution (80% ethanol), and standard curves of 0–0.03 mg/L 4’-O-methyl-gallocatechin, and 0–0.16 g/L rPAC were utilized as references.

Geraniin quantification

Geraniin was quantified using a modification of a previously reported method [17]. An LC20 HPLC-DAD chromatographer (from Shimadzu Corp., Kyoto, Japan) equipped with a Dionex Acclaim 120, C18-column (150 mm 4,6 mm i.d., 5 µm) was used to quantify the amount of geraniin. 20µL of the sample was utilized. The mobile phase consisted of (A) 0,1% formic acid and (B) acetonitrile with a gradient from 0-28 min, 0-20% B in A, 28-34 min, and 20-70% B in A. The flow rate was 0,8 mL/min. The detection was performed using 280 nm.

Determination of Antimicrobial Activity

Antibacterial activity was determined using the Kirby-Bauer disc diffusion method as previously reported [18]. Each disc was impregnated with 50 µL of each of the extract. 30 µg/mL chloramphenicol has been used as the positive control, and water:acetone (7:3) as the negative control. *Escherichia coli ATCC 25922, Pseudomonas aeruginosa ATCC 9027, Staphylococcus aureus ATCC 25923*, and *Bacillus subtilis ATCC 6633* were utilized to test susceptibility.

Results and discussion

Optimization of the extraction procedure

Figure 1(A) shows the results of the solvent choice to extract rambutan shells. All the solvents tested were enriched with HCl 5% to increase their stability. Rambutan extracts are stable at a pH lower than 3 [19]. Acetone, short alcohols, and water were components included in the potential extraction solvents tested. A significant difference is observed in ethanol 95% over the other solvent mixtures.

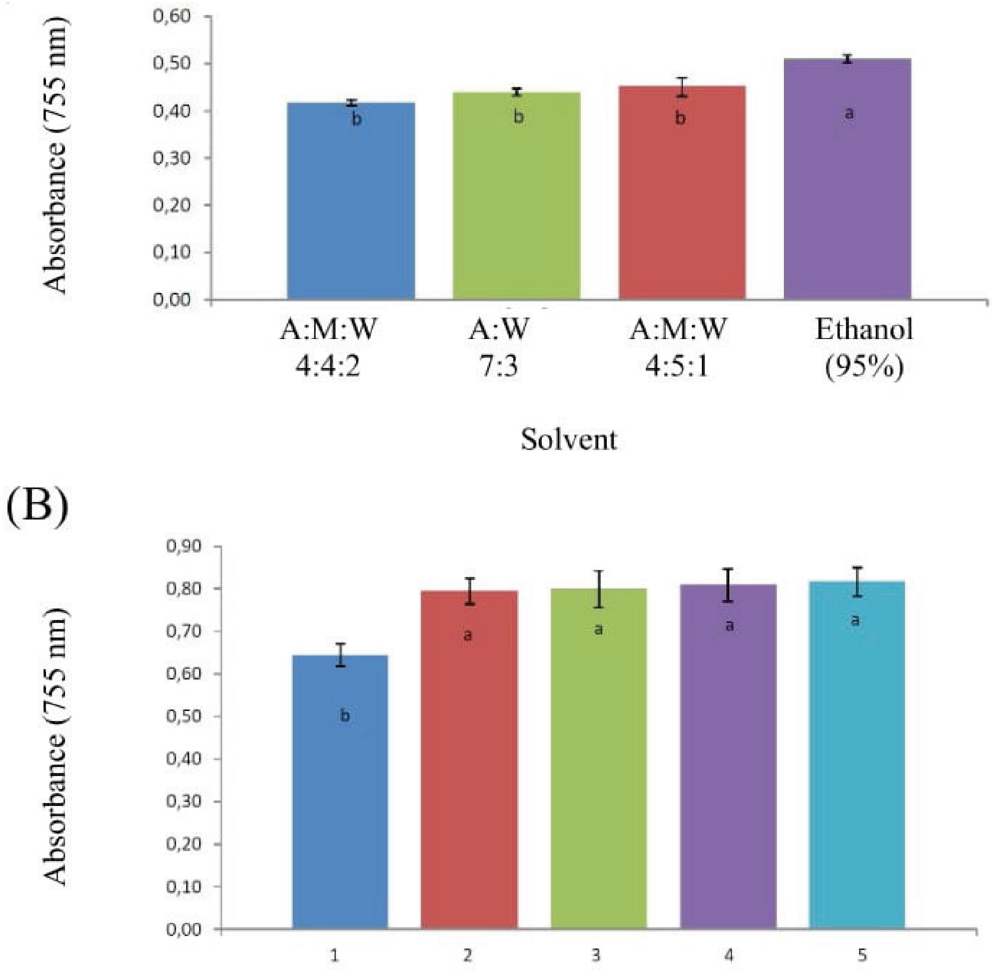


Figure 1. Optimal extraction conditions. (A) Solvent selection. A: Acetone, M: methanol, and W: water. All solvents contain 5% HCl (B) Number of extraction cycles. Compact display letters represent Tukey´s test. α=0.05, n=3.

Figure 1(B) also represents the number of extraction cycles. Every cycle represents 2 mL of solvent, sonication, and centrifugation. The sonication step helps to destroy cell structures and accelerates the extraction process. This method is fast and utilizes small solvent volumes. According to Das [20], ultrasound-assisted extraction efficiency is slightly lower than maceration or Soxhlet extraction. However, several cycles usually help to increase efficiency to a similar level. Also, thermal treatment, such as a Soxhlet can induce the degradation

of thermolabile metabolites, The Rambutan extraction process shows no significant differences after the second cycle. The metabolites extraction process is very fast and efficient for rambutan, respecting other matrices. E.g., 3 extraction cycles are required for both corn (*Zea mays*) [14] and ber (*Ziziphus mauritiana*) [21].

TPC and antioxidant activity of crude extracts

Previously, some compounds have been identified in the literature as main components for rambutan extracts [22, 23] geraniin, corilagin, rutin, ellagic acid, gallic acid, and some derivatives. Folin-Ciocalteu´s test for TPC reacts with phenolic substances (and some other reducing compounds), while the antioxidant tests are based on two different mechanisms: hydrogen atom transfer (such as in the ORAC method), and electron transfer (such as in the DPPH method) [24]. Phenolic compounds are considered the main contributors to the antioxidant capacity of rambutan extracts [22]. Figures 2(C) and 2(F) show the relationship between ORAC vs TPC and DPPH vs TPC. Pearson coefficients are 0.6981, and 0.5240, respectively. The correlations are low compared to other types of samples such as some teas, where the Pearson coefficient takes values ranging from 0.99 to 1.00 [24]. Individual contributions of rambutan components have been assessed in the past and found significantly different. Geraniin, corilargin, and ellagic acid have 32-65% antioxidant capacity of gallic acid [23]. Then, low correlations can be explained by different distributions of phenolic compounds.

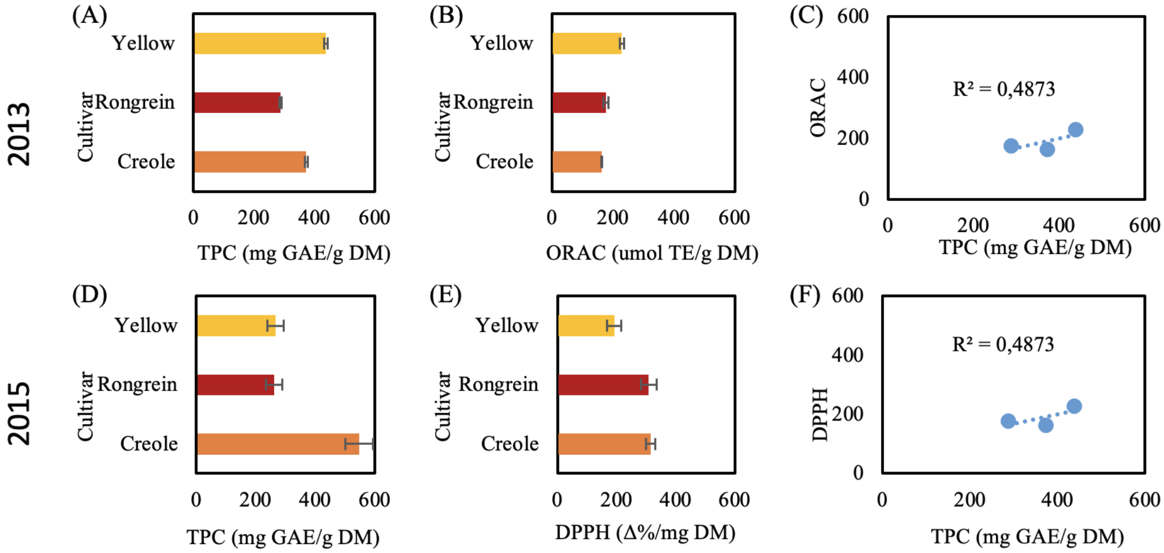


Figure 2. Total Phenolic Compounds (TPC), antioxidants (DPPH and ORAC), and Proanthocyanidins (PAC) from selected rambutan samples.

Samples harvested during 2013 and 2015 (Figures 2(A) and 2(D)) have different TPC content. Yellow variety showed the highest from the first group (438 mg GAE/g) but decreased to 266 mg GAE/g in 2015. The Creole variety was the second highest during the 2013 harvest (373 mgGAE/g) and then increased to 547 mgGAE/g, which is the highest value. Rongrein variety just suffers small variations in TPC concentration, being the middle value for both sets of samples (288 and 261 mgGAE/g).

Proanthocyanidins-enriched extracts

The rPAC fraction is prepared by purifying PACs from the crude extract. This procedure leads to a “self-fruit” standard, in a similar way to previous reports [25]. rPAC standard is more representative of the heterogenicity of PACs present in rambutan than a commercial reference such as 4-MGC. In this work, we quantified TPC content, DPPH antioxidant activity, and PACs content from the rPAC fraction extracted from the three varieties included in this study. Results are shown in Figure 3.

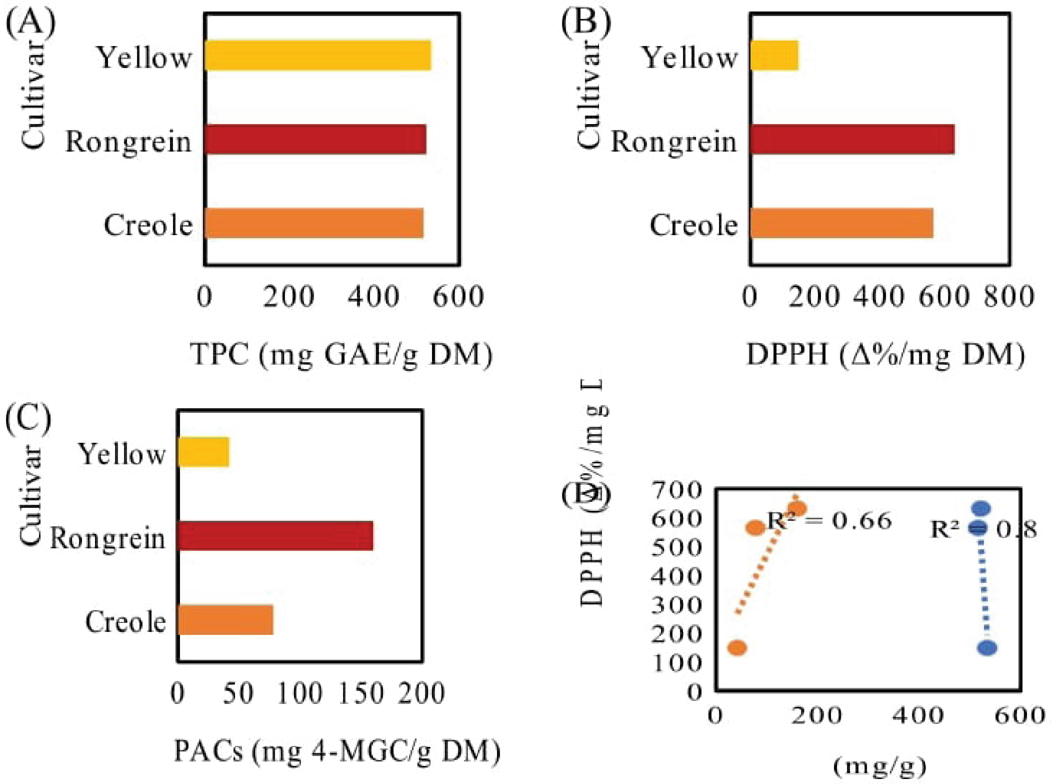


Figure 3. Total phenolic compounds (TPC), antioxidants (DPPH), and proanthocyanidins (PAC) from rPACs.

TPC of the rPACs of the three varieties ranges from 516-534 mg GAE/g. Values are very consistent probably because rPACs are supposed to be near 100% proanthocyanidins. Flavonoids such as PACs and other phenolic compounds are the most abundant pigments in fruits, including rambutan peels [26]. Then, more abundant conjugated, or delocalized aromatic systems can be found in red varieties (such as rongrein and creole) than in the yellow varieties.

Figure 3(D) shows the Pearson coefficient for DPPH vs TPC or PACs. The correlation between PACs and antioxidant activity is clearer than in the crude extracts. The lack of other contaminants (because of the purification of rPACs) seems to help to establish a better relationship. However, PACs from the three varieties are different in chemical nature.

Geraniin analysis

Some ellagitannins have particular attention because they are reported to have several bioactivities. Geraniin, an ellagitannin previously reported in *N. lappaceum* skins has been selected to be quantified. Figure 4 shows the composition of the three varieties included in this study.

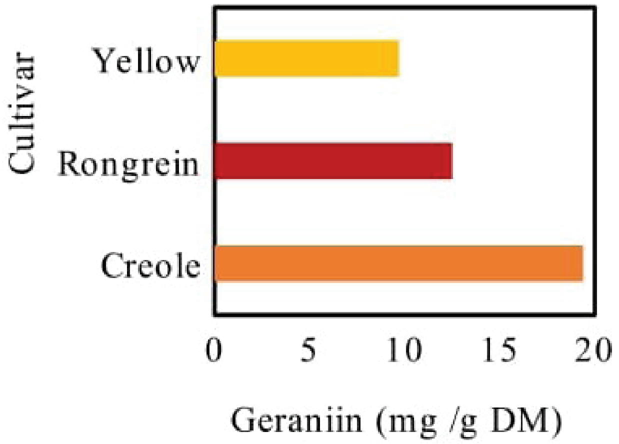


Figure 4. Geraniin composition of a sample of three rambutan varieties.

Concentrations of geraniin range from 9.7 to 19.4 mg/g. Creole variety reached the highest concentration.

Antimicrobial activity

Figure 5 shows the results for the antimicrobial activities of rambutan extracts against four bacterial strains.

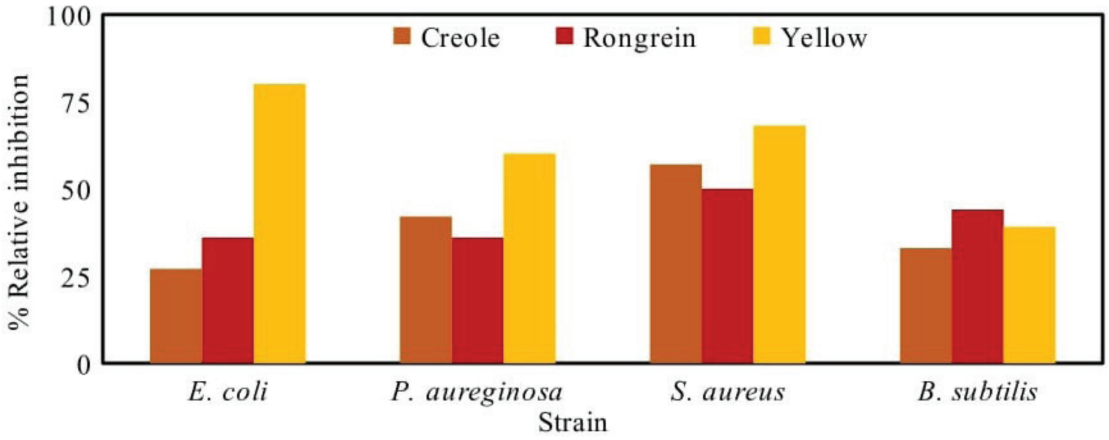


Figure 5. Antimicrobial activity of 50 µL of rambutan extracts. 30 µg chloramphenicol has been used as positive control, and water:acetone (7:3) as negative control. n=3. Error bars represent standard deviations. Letters on top of columns represent Tukey test grouping (samples containing the same letter belong to the same group), per bacteria tested.

The three varieties have antibacterial activity against the four microorganisms tested. Bacteria included two gram-positive and two gram-negative microorganisms. Also, the yellow variety of rambutan showed the highest antimicrobial activity, reaching 80% relative inhibition for *E. coli* and 60 and 68% for *P. aeruginosa* and *S. aureus,* respectively. These results are similar to those found in other publications [3, 23], in which the antibiotic activity is related to the phenolic content of the extracts.

Conclusions

The three varieties of rambutan included in this study (Creole, Yellow, and Rongrein) are antioxidant-rich in their peels. The concentration between varieties does not keep a pattern between the samples of two different years included in this study. TPC composition is 250-550 mg GAE/g in all the samples, and 10-20 mg/g geraniin. All the extracts have shown antimicrobial properties and antioxidant activity.

Peels from Yellow, Rongrain and Creole rambutan are suitable for both human and animal supplementation according to the chemical composition.

Acknowledgments

Authors thank Eng. Melvin Arias-Tenorio, from the Ministry of Agriculture and Livestock of Costa Rica, for the identification of the rambutan varieties. Also, the local producers Eduardo Delgado and Luis Zuñiga for providing the material.

References

[1] C. Hernández-Hernández *et al.*, "Rambutan (Nephelium lappaceum L.): Nutritional and functional properties," *Trends in food science technology,* vol. 85, pp. 201-210, 2019.

[2] Z. Tingting, Z. Xiuli, W. Kun, S. Liping, and Z. Yongliang, "A review: extraction, phytochemicals, and biological activities of rambutan (Nephelium lappaceum L) peel extract," *Heliyon,* 2022.

[3] N. N. M. Phuong, T. T. Le, J. Van Camp, and K. Raes, "Evaluation of antimicrobial activity of rambutan (Nephelium lappaceum L.) peel extracts," *International journal of food microbiology,* vol. 321, p. 108539, 2020.

[4] A. Perera, S. H. Ton, and U. D. Palanisamy, "Perspectives on geraniin, a multifunctional natural bioactive compound," *Trends in Food Science Technology,* vol. 44, no. 2, pp. 243-257, 2015.

[5] H. S. Cheng, S. H. Ton, and K. Abdul Kadir, "Ellagitannin geraniin: a review of the natural sources, biosynthesis, pharmacokinetics and biological effects," *Phytochemistry reviews,* vol. 16, pp. 159-193, 2017.

[6] K. Mahmood, H. Kamilah, A. K. Alias, and F. Ariffin, "Nutritional and therapeutic potentials of rambutan fruit (Nephelium lappaceum L.) and the by-products: a review," *Journal of Food Measurement Characterization,* vol. 12, pp. 1556-1571, 2018.

[7] H. Deng *et al.*, "Comparative nutritional and metabolic analysis reveals the taste variations during yellow rambutan fruit maturation," *Food Chemistry: X,* vol. 17, p. 100580, 2023.

[8] T. Manso, M. Lores, and T. de Miguel, "Antimicrobial activity of polyphenols and natural polyphenolic extracts on clinical isolates," *Antibiotics,* vol. 11, no. 1, p. 46, 2021.

[9] D. Yun, Y. Qin, J. Zhang, M. Zhang, C. Qian, and J. Liu, "Development of chitosan films incorporated with rambutan (Nephelium lappaceum L.) peel extract and their application in pork preservation," *International Journal of Biological Macromolecules,* vol. 189, pp. 900-909, 2021.

[10] U. Sukatta *et al.*, "Rambutan (Nephelium lappaceum) peel extract: Antimicrobial and antioxidant activities and its application as a bioactive compound in whey protein isolate film," *Songklanakarin Journal of Science Technology,* vol. 43, no. 1, 2021.

[11] T. Ampapon and M. Wanapat, "Dietary rambutan peel powder as a rumen modifier in beef cattle," *Asian-Australasian Journal of Animal Sciences,* vol. 33, no. 5, p. 763, 2020.

[12] C. Le Xuan *et al.*, "Dietary inclusion of rambutan (Nephelium lappaceum L.) seed to Nile tilapia (Oreochromis niloticus) reared in biofloc system: Impacts on growth, immunity, and immune-antioxidant gene expression," *Fish Shellfish Immunology,* vol. 122, pp. 215-224, 2022.

[13] C. Le Xuan *et al.*, "Effects of dietary rambutan (Nephelium lappaceum L.) peel powder on growth performance, immune response and immune-related gene expressions of striped catfish (Pangasianodon hypophthalmus) raised in biofloc system," *Fish Shellfish Immunology,* vol. 124, pp. 134-141, 2022.

[14] R. Syedd-León, R. Orozco, V. Álvarez, Y. Carvajal, and G. Rodríguez, "Chemical and antioxidant charaterization of native corn germplasm from two regions of costa rica: A conservation approach," *International journal of food science,* vol. 2020, 2020.

[15] V. Bondet, W. Brand-Williams, and C. Berset, "Kinetics and mechanisms of antioxidant activity using the DPPH. free radical method," *LWT-Food Science Technology,* vol. 30, no. 6, pp. 609-615, 1997.

[16] P. Brescia, "Determination of Antioxidant potential using an Oxygen Radical Absorbance Capacity (ORAC) Assay with Synergy TM H4," *BioTek Application Note,* pp. 4-12, 2012.

[17] A. Tuominen and T. J. P. A. Sundman, "Stability and oxidation products of hydrolysable tannins in basic conditions detected by HPLC/DAD–ESI/QTOF/MS," vol. 24, no. 5, pp. 424-435, 2013.

[18] B. Vega-López *et al.*, "Phytonutraceutical evaluation of five varieties of tomato (Solanum lycopersicum) during ripening and processing," *LWT,* p. 113592, 2022.

[19] J. Sun, H. Peng, W. Su, J. Yao, X. Long, and J. Wang, "Anthocyanins extracted from rambutan (Nephelium lappaceum L.) pericarp tissues as potential natural antioxidants," *Journal of Food Biochemistry,* vol. 35, no. 5, pp. 1461-1467, 2011.

[20] S. Das, A. Ray, N. Nasim, S. Nayak, and S. Mohanty, "Effect of different extraction techniques on total phenolic and flavonoid contents, and antioxidant activity of betelvine and quantification of its phenolic constituents by validated HPTLC method," *Biotech,* vol. 9, no. 1, p. 37, 2019.

[21] E. Cubero-Román, Y. Carvajal-Miranda, G. Rodríguez, V. Álvarez-Valverde, and P. Jiménez-Bonilla, "Antioxidant and antimicrobial activity of two Costa Rican cultivars of ber (Ziziphus mauritiana): An underexploited crop in the American tropic," *Food Science and Nutrition* vol. 11, no. 6, pp. 3320-3328, 2023.

[22] N. N. M. Phuong, T. T. Le, M. Q. Dang, J. Van Camp, and K. Raes, "Selection of extraction conditions of phenolic compounds from rambutan (Nephelium lappaceum L.) peel," *Food Bioproducts Processing,* vol. 122, pp. 222-229, 2020.

[23] N. Thitilertdecha, A. Teerawutgulrag, J. D. Kilburn, and N. Rakariyatham, "Identification of major phenolic compounds from Nephelium lappaceum L. and their antioxidant activities," *Molecules,* vol. 15, no. 3, pp. 1453-1465, 2010.

[24] M. K. Roy, M. Koide, T. P. Rao, T. Okubo, Y. Ogasawara, and L. R. Juneja, "ORAC and DPPH assay comparison to assess antioxidant capacity of tea infusions: Relationship between total polyphenol and individual catechin content," *International journal of food sciences nutrition,* vol. 61, no. 2, pp. 109-124, 2010.

[25] E. R. Gullickson, C. G. Krueger, A. Birmingham, M. Maranan, and J. D. Reed, "Development of a cranberry standard for quantification of insoluble cranberry (Vaccinium macrocarpon Ait.) proanthocyanidins," *Journal of agricultural food chemistry,* vol. 68, no. 10, pp. 2900-2905, 2019.

[26] Q. Qi *et al.*, "Anthocyanins and proanthocyanidins: Chemical structures, food sources, bioactivities, and product development," *Food Reviews International,* vol. 39, no. 7, pp. 4581-4609, 2023.

Evaluation of the antimicrobial properties of natural extracts of *Ganoderma lucidum*

Evaluación de las propiedades antimicrobianas de extractos naturales de *Ganoderma lucidum*

Karina Abarca-Cascante[1], Nicole Arias-Espinoza[2], Judith Cambronero-Vega[3], Bryan Zúñiga-Gaitán[4], Victor Álvarez-Valverde[5], José B. Azofeifa-Bolaños[6], Jorengeth Abad Rodríguez-Rodríguez[7]

Abarca-Cascante, K; Arias-Espinoza, N; Cambronero-Vega, J; Zúñiga-Gaitán, B; Álvarez-Valverde, V; Azofeifa-Bolaños, J. B; Rodríguez-Rodríguez, J. A. Evaluation of the antimicrobial properties of natural extracts of *Ganoderma lucidum*. *Tecnología en Marcha*. Vol. 37, special issue. August, 2024. IEEE International Conference on BioInspired Processing. Pag. 68-74.

https://doi.org/10.18845/tm.v37i7.7301

[1] Escuela de Ciencias Biológicas, Universidad Nacional. Costa Rica.
karina.abarca.cascante@est.una.ac.cr
https://orcid.org/0009-0006-8633-4678

[2] Escuela de Ciencias Biológicas, Universidad Nacional. Costa Rica.
nicole.arias.espinoza@est.una.ac.cr
https://orcid.org/0000-0002-0288-3586

[3] Escuela de Ciencias Biológicas, Universidad Nacional. Costa Rica.
judith.cambronero.vega@est.una.ac.cr
https://orcid.org/0000-0003-0570-6400

[4] Escuela de Ciencias Biológicas, Universidad Nacional. Costa Rica.
bryan.zuniga.gaitan@est.una.ac.cr
https://orcid.org/0000-0003-2259-608X

[5] Laboratorio de Fitoquímica (LAFIT), Escuela de Química, Universidad Nacional, Heredia. Costa Rica.
victor.alvarez.valverde@una.cr
https://orcid.org/0000-0001-6007-9150

[6] Laboratorio de Biotecnología Plantas, Escuela de Ciencias Biológicas, Universidad Nacional. Costa Rica.
bernal.azofeifa.bolanos@una.cr
https://orcid.org/0000-0002-8902-0352

[7] Laboratorio de Biotecnología Microbiana, Escuela de Ciencias Biológicas, Universidad Nacional. Costa Rica.
jorengeth.rodriguez.rodriguez@una.cr
https://orcid.org/0000-0001-8452-8256

Keywords

Ethanol extract; ganoderic acid; triterpenoids; polysaccharides; bioactive compounds.

Abstract

There is a growing importance of alternative anticancer and anti-inflammatory treatments. Current interest lies in natural medicine and secondary metabolites present in plants, particularly focusing on triterpenes. *Ganoderma lucidum* is a fungal species of significant commercial interest due to its therapeutic and medicinal properties resulting from secondary metabolites such as polysaccharides and triterpenes. These compounds exhibit antitumor activity and bolster the immune system. The study aimed to assess the antimicrobial activity of *G. lucidum* extracts, both commercial and wild. The content of ganoderic acids in all extracts showed a slight difference in concentration between commercial and wild extracts. However, no bacterial inhibition halos were observed in any of the strains used. The presence of varying concentrations of ganoderic acids among treatments underscores the importance of optimizing and standardizing a comprehensive strategy for extracting secondary metabolites, focusing on producing high-quality supplements and pharmaceutical products. Furthermore, preserving the stability of the obtained triterpenes is necessary due to their importance in the medicinal properties of the fungus.

Palabras clave

Extracto etanólico; ácido ganodérico; triterpenoides; polisacáridos; compuestos bioactivos.

Resumen

Existe una creciente importancia de los tratamientos alternativos anticancerígenos y antiinflamatorios. El interés actual radica en la medicina natural y los metabolitos secundarios presentes en plantas, con un enfoque particular en los triterpenos. *Ganoderma lucidum* es una especie fúngica de gran interés comercial por sus propiedades terapéuticas y medicinales debido a la presencia de metabolitos secundarios como polisacáridos y triterpenos, que presentan actividad antimicrobiana. El objetivo del estudio fue la evaluación de la actividad antimicrobiana de extractos de *G. lucidum*, tanto comerciales como silvestres. El contenido de ácidos ganodéricos en todos los extractos tienen una leve diferencia en la concentración entre los extractos comerciales y silvestres. Sin embargo, no se obtuvieron halos de inhibición bacteriana en ninguna de las cepas utilizadas. La presencia de diferentes concentraciones de ácidos ganodéricos entre los tratamientos, resalta la importancia de optimizar y estandarizar una estrategia integral de extracción de metabolitos secundarios, con un enfoque dirigido hacia la producción de suplementos y productos farmacéuticos de alta calidad. Además, es necesario preservar la estabilidad de los triterpenos obtenidos, dada su importancia en las propiedades medicinales del hongo.

Introduction

The fungus *Ganoderma lucidum* (commonly known as Reishi) is naturally distributed in Central America, characterized by a basidiocarp that grows near tree trunks and is renowned in medicinal fields for treating diseases. Polysaccharides and triterpenes are their primary components with significant physiological activity [1]. There exists substantial potential as an antiviral agent due to its capacity to enhance immune defenses through ganoderic acids (GA), ganodermanondiol, lucidumol, and ganodermanontriol. Commonly employed extraction methods include hot water extraction, ultrasonic bath, maceration, and solvent-based extraction [2]. Quantification of different metabolites is primarily conducted using HPTLC techniques [3].

The key secondary metabolites of *G. lucidum* are triterpenes, which belong to the subgroup of terpenes [4]. These exhibit antimicrobial activity by disrupting specific sites on the bacterial plasma membrane [5]. Therefore, this study aimed to evaluate the antimicrobial activity of both commercial and wild natural extracts of *G. lucidum,* in addition to obtaining *Ganoderma* profiles (GA) for their potential use in the pharmaceutical industry.

Materials and Methods

Collection of Fungi

The wild mushrooms were collected from the green areas of the Omar Dengo Campus of the National University, located in Heredia, Costa Rica (latitude 10°00 ‘02.08” N, longitude 84°06 ‘34.61” W) in September 2023. On the other hand, the commercial powdered extract was obtained from a Costa Rican distributor. The experiments were conducted at the Teaching Biotechnology Laboratory (LABID), the Phytochemistry Laboratory (LAFIT), and the Microbial Biotechnology Laboratory (LABIMI) of the same university.

Extraction Process

The fungus was lyophilized, ground, and sieved to obtain particles smaller than 1 mm [7]. On the other hand, for the extract preparation, 2 g of mushroom and 200 mg of the commercial sample were weighed and dissolved in 15 mL and 2 mL of 95% ethanol:water (7:3), respectively [3]. The wild extract was obtained through sonication in an ultrasonic bath (40 KHz) for 5 minutes. The supernatant obtained was passed through a 0.45 µm filter using a syringe. The process was carried out in triplicate until a volume of 15 mL was obtained. Subsequently, the *G. lucidum* extract was concentrated using the Büchi Rotavapor R-200 equipment with a water bath at 40°C [8]. Once the solvent was evaporated, the solids were weighed and dissolved in 95% ethanol and water at a 1:1 ratio to achieve a final concentration of 100 mg/mL for both extracts.

Determination of terpenes by HPTLC

The Camag Autosampler ATS4, Derivatizer, and CAMAG TLC Visualizer 3 equipment were employed. The selected mobile phase consisted of a mixture of dichloromethane:ethyl acetate:cyclohexane: formic acid:ethanol (8:3:9:0.8:0.5) [8]. Aluminum-backed plates with a stationary phase composed of Silica gel 60 F254, measuring 20 x 10 cm, were used in the procedure. For band visualization, the plates were exposed to wavelengths of 366 nm and 254 nm. Additionally, the following reference standards were incorporated: ganoderic acid AG-B, AG-D, and AG-F from the commercial house Sigma Aldrich, with the purpose of identifying the secondary metabolites of interest.

Antimicrobial assessment

The fungal extract was used to evaluate its antimicrobial inhibition capacity by forming inhibition zones. To achieve this, bacterial strains of *Escherichia coli*, *Staphylococcus aureus* and *Proteus vulgaris* were cultured with a concentration of 0.5 on the McFarland scale. These strains were transferred to test tubes containing 0.85% saline solution. Subsequently, 50 mL Falcon tubes were prepared with 25 mL of Muller-Hinton agar and 1 mL of each bacterial strain. After pouring the agar into Petri dishes, wells were created for the *Ganoderma* extracts and Streptomycin/Penicillin (500 µg/mL) as a positive control, dispensing 50 µL in the wells. Finally, the Petri dishes were incubated for 24 hours at 35°C for observation and measurement of the inhibition zones [5]. The collected data underwent statistical analysis using a mixed linear model with Poisson distribution, specifically designed to address the problem of inflated zero counts [6], and these were analyzed using the R Studio program (version 4.2.2).

Results

In Figure 1, the HPTLC chromatography for GA is shown, confirming the presence of terpenoids in all extracts obtained from *G. lucidum*. The band of GA from the wild ethanol extract (GA Wild *Ganoderma*) is observed with less intensity compared to the bands of the commercial extract. However, the concentrated sample of the wild extract (GA Wild *Ganoderma* 5:1) displays improved GA bands and a greater variety of compounds.

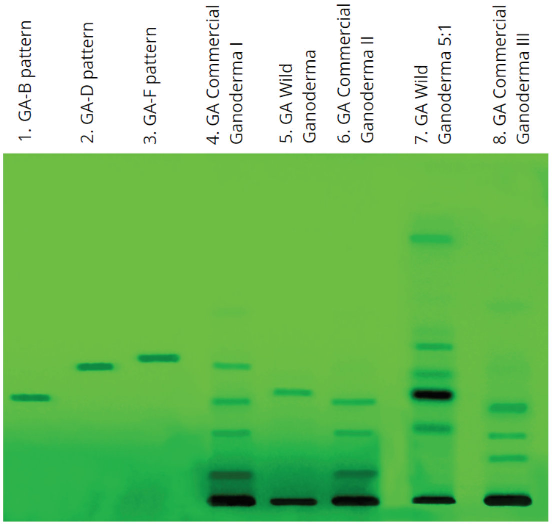


Figure 1. Profiles of ganoderic acid were analyzed from commercial and wild *G. lucidum* extracts using HPTLC. Band profile: Lane 1, GA-B standard. Lane 2, GA-D. Lane 3, GA-F. Lane 4, Commercial *Ganoderma* 1. Lane 5, Unconcentrated Wild *Ganoderma*. Lane 6, Duplicate Commercial *Ganoderma*. Lane 7, Concentrated Wild *Ganoderma*. Lane 8, Triplicate Commercial *Ganoderma*

In the evaluation of the antimicrobial effect, no bacterial inhibition halos were detected in any of the strains used for both extracts (F:17.45; df:2; $p<0.05$), despite the presence of ganoderic acids in the samples (see Figure 2).

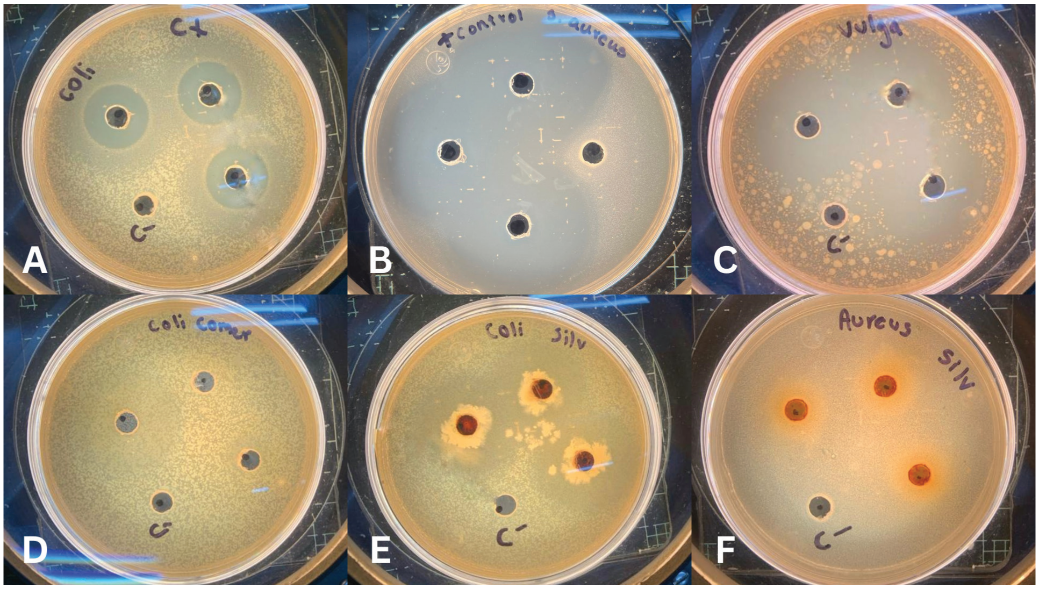


Figure 2. Antimicrobial tests of ethanolic extracts. Three positive controls with antibiotics are presented for three different strains: (A) *E. coli*, (B) *S. aureus*, and (C) *P. vulgaris*. In addition, plates with (D) commercial extract in *E. coli*, (E) wild extract in *E. coli* and (F) wild extract in *S. aureus* are shown, with a comparison to the positive controls

Discussion

In the chromatographic analysis, the absence of GA-F in the extracts indicates a lack of similarity with the reference pattern [9]. These differences are related to the generation of GA, specifically in the lanosterol biosynthesis process, which is influenced by structural changes through acetylation, methylation, and hydroxylation reactions [10]. On the other hand, the proper composition of GA is influenced by the substrate in which the fruiting body is found, due to the presence of fundamental enzymes that catalyze triterpene differentiation reactions, which are specific to certain substrates [11].

In the antibacterial tests conducted, the lack of bacterial inhibition could be related to the solvent used or the extraction time [7, 12]. Another potential extract type is the acetonic extracts, which have been reported to possess antimicrobial activity [13]. Additionally, the contents of antimicrobial compounds vary according to the culture conditions as well as the growth stage of *G. lucidum* [14]. In this case, to obtain high-quality extracts, it is recommended to employ specific culture media and adjust the culture conditions to maximize the yield of *G. lucidum* [15].

The stability of terpenes is of utmost importance due to their fundamental role in the medicinal properties of the fungus. The degradation of triterpenoids can compromise the quality and efficacy of the extracts. It is noteworthy that many triterpenes have demonstrated a wide range of biological activities, emphasizing the need to preserve their stability for application [16].

The need for optimization and standardization of bioactive extraction significantly impacts obtaining high-quality extracts and has the potential to drive the production of high-quality supplements and pharmaceutical products. This optimization is crucial for both public health benefits and the country's economic development [17]. Despite the recent disclosure of data related to the genome and transcriptome of *G. lucidum*, there are still areas of knowledge that need exploration, including its potential in the antimicrobial field [18].

Conclusion

There was no antimicrobial activity in both the commercial and wild extracts, as they did not reveal inhibitory effects on bacterial growth. It is suggested that factors such as time and the extraction solvent were not suitable for the specific growth conditions of the fungus. Finally, the optimization and standardization of extraction methods allow for producing high-quality extracts with potential applications in the pharmaceutical industry.

Acknowledgments

We would like to thank our colleague Andrew Lindwedel Cruz for the guidance provided during the analysis of the results.

References

[1] L. F. Arce-Torres, I. Gómez-Díaz, M. Monge-Artavia, J. Prado-Cordero, “Metabolitos secundarios con actividad medicinal extraídos de hongos provenientes de Centroamérica”, *Revista Tecnología En Marcha*, vol. 33, no. 3, pp. 80–89, 2020. https://doi.org/10.18845/tm.v33i3.4416

[2] E. Ekiz, E. Oz, A. M. A. El-Aty, C. Proestos, C. S. Brennan, M. Zeng, I. Tomasevic, T. Elobeid, K. Çadirci, M. Bayrak, & F. Oz, “Exploring the potential medicinal benefits of *Ganoderma lucidum*: From metabolic disorders to coronavirus infections,” *Foods*, vol. 12, no. 7, p. 1512, 2023. https://doi.org/10.3390/foods12071512

[3] S. Zheng, W. Zhang, & S. Liu, “Optimization of ultrasonic-assisted extraction of polysaccharides and triterpenoids from the medicinal mushroom *Ganoderma lucidum* and evaluation of their in vitro antioxidant capacities,” *PLOS ONE*, vol. 15, no. 12, e0244749, 2020. https://doi.org/10.1371/journal.pone.0244749

[4] R. A. Hill & J. D. Connolly, “Triterpenoids,” *Natural Product Reports*, vol. 37, no. 7, pp. 962–998, 2019. https://doi.org/10.1039/c9np00067d

[5] L. A. Segovia-Tello, “Evaluación de la actividad antimicrobiana del extracto etanólico obtenido de Ganoderma lucidum frente a *Escherichia coli*, *Staphylococcus aureus* y *Proteus mirabilis,*” Tesis de Grado, Escuela Superior Politécnica de Chimborazo, 2017.

[6] Hilbe, J. M., “The statistical analysis of count data/El análisis estadístico de los datos de recuento”, *Cultura y Educación,* vol. 29, no. 3, 409-460., 2017

[7] R. Noverita & Y. H. LumbanTobing, “Antibacterial Activities of Ethanol Extracts Fruit Bodies of *Ganoderma lucidum* and *Rigidoporus microphorus* Against *Escherichia coli* and *Staphlyococcus aureus*,” *Journal of Tropical Biodiversity,* vol. 1, no. 1, pp. 35-46, 2020. http://e-journal.unas.ac.id/index.php/bio/article/view/25

[8] M. Abate, G. Pepe, R. Randino, S. Pisanti, M. G. Basilicata, V. Covelli, M. Bifulco, W. Cabri, A. M. D’Ursi, P. Campiglia, & M. Rodriquez, “*Ganoderma lucidum* ethanol extracts enhance Re-Epithelialization and prevent keratinocytes from Free-Radical injury,” *Pharmaceuticals,* vol. 13, no. 9, p. 224, 2020. https://doi.org/10.3390/ph13090224

[9] D. Frommenwiler, D. Trefzer, M. Schmid, S. Cañigueral, & E. Reich, “Comprehensive HPTLC Fingerprinting: A novel economic approach to evaluating the quality of *Ganoderma lucidum* fruiting body,” *Journal of Liquid Chromatography & Related Technologies*, vol. 43, no. 11-12, pp. 414-423, 2020. https://doi.org/10.1080/10826076.2020.1725560

[10] W. Wang, H. Xiao, & J. Zhong, “Biosynthesis of a novel ganoderic acid by expressing CYP genes from *Ganoderma lucidum* in *Saccharomyces cerevisiae,*” *Applied Microbiology and Biotechnology*, vol. 106, no. 2, pp. 523-534, 2021. https://doi.org/10.1007/s00253-021-11717-w

[11] P. Bondzie-Quaye, M. S. Swallah, A. Acheampong, S. M. Elsherbiny, E. O. Acheampong, & Q. Huang, “Advances in the biosynthesis, diversification, and hyperproduction of ganoderic acids in *Ganoderma lucidum,*” *Mycological Progress*, vol. 22, no. 4, pp. 1-17, 2023. https://doi.org/10.1007/s11557-023-01881-w

[12] S. Quereshi, A. Pandey, & S. Sandhu, “Evaluation of antibacterial activity of different *Ganoderma lucidum* extract,” *People’s Journal of Scientific Research*, vol. 3, no. 1, pp. 9-13, 2010. https://doi.org/10.5281/zenodo.8285679

[13] D. Cör, Ž. Knez, & L. Chen, “Antitumour, antimicrobial, antioxidant and antiacetylcholinesterase effect of *Ganoderma lucidum* terpenoids and polysaccharides: a review,” *Molecules*, vol. 23, no. 3, p. 649, 2018. https://doi.org/10.3390/molecules23030649

[14] C. Rodríguez-Farinango, J. Pineda-Insuasti, D. A. B. Revelo, F. Ariel, M. Puetate, & C. A. P. Soto, "Producción de *Ganoderma lucidum* y su potencial medicinal: una revisión," *Revista Biorrefinería*, vol. 4, no. 4, pp. 1–9, 2021. https://www.cebaecuador.org/wp-content/uploads/2022/01/17.pdf

[15] A. M. Torres López, J. C. Quintero Díaz, & L. Atehortua Garcés, "Efecto de nutrientes sobre la producción de biomasa del hongo medicinal *Ganoderma lucidum*," *Revista Colombiana de Biotecnología*, vol. 13, no. 1, pp. 103-109, 2011. https://doaj.org/article/8e9e094bf8ab4827a9b89bd4a221cc92

[16] D. Castañeda-Antonio, "Caracterización fisicoquímica del extracto estandarizado del hongo medicinal *Ganoderma lucidum* y análisis de su impacto potencial en la salud humana," Tesis doctoral en ciencias, Instituto de enseñanza e investigación en ciencias agrícolas, 2019.

[17] R. A. Alsaheb, K. Z. Zjeh, R. A. Malek, J. K. Abdullah, A. El Baz, N. El Deeb, et al., "Bioprocess optimization for exopolysaccharides production by *Ganoderma lucidum* in semi-industrial scale," *Recent Patents on Food, Nutrition & Agriculture*, vol. 11, no. 3, pp. 211-218, 2020. https://doi.org/10.2174/2212798411666200316153148

[18] G. J. Yu, Y. L. Yin, W. H. Yu, W. Liu, Y. X. Jin, A. Shrestha, et al., "Proteome exploration to provide a resource for the investigation of *Ganoderma lucidum*," *PLoS One*, vol. 10, no. 3, e0119439, 2015. https://doi.org/10.1371/journal.pone.0119439

Clustering of cantons in Costa Rica based on interest variables during the beta variant of Covid-19

Agrupamiento de los cantones de Costa Rica con base en variables de interés durante la variante beta del Covid-19

Isaí Ugalde-Araya[1]

Ugalde-Araya, I. Clustering of cantons in Costa Rica based on interest variables during the beta variant of Covid-19. *Tecnología en Marcha*. Vol. 37, special issue. August, 2024. IEEE International Conference on BioInspired Processing. Pag. 75-80.

https://doi.org/10.18845/tm.v37i7.7302

[1] Statistician and Researcher. Advanced Computing Laboratory. National High Technology Center. Costa Rica.
iugalde@cenat.ac.cr
https://orcid.org/0009-0003-9517-1653

Keywords

Cluster analysis; mortality rate; morbidity rate; case fatality rate; third wave pandemic; distance.

Abstract

The study of case behavior and the analysis of bioindicators are relevant and important for decision-making by health authorities worldwide, related to the Covid-19 pandemic. Thus, numerous investigations have been carried out around the world to understand this phenomenon, its variants, and its primary impacts on population health. In this study, a cluster analysis was conducted based on the variables of mortality rate, morbidity rate, and fatality rate, along with cantonal geographical density, for the period of the beta variant in Costa Rica, corresponding to the months from February to June 2021. Therefore, a total of three methods were chosen to obtain groups: k-means, k-medoids, and fuzzy methods; as well as two types of distances: Euclidean and Manhattan. Additionally, the sum of squares within groups and the Dunn index were used to validate the formation of the clusters. It was identified that the method and distance that formed the most compact cantonal clusters with lower intragroup variability were k-medoids and Manhattan, respectively, due to their greater robustness against extreme values. Among the formed groups, cluster 1 has a moderate impact of the pandemic during the specified variant, while groups 2 and 3 have low and high impacts, respectively. Moreover, groups 1 and 2 are predominantly composed of cantons outside the Greater Metropolitan Area, in contrast to the third group. This analysis provides valuable insights for health authorities in understanding the impacts of the Covid-19 pandemic in Costa Rican regions and aids in the development of targeted strategies for effective management.

Palabras clave

Análisis de conglomerados; tasa de mortalidad; tasa de morbilidad; tasa de letalidad; tercera ola pandémica; distancia.

Resumen

El estudio del comportamiento de los casos y el análisis de bioindicadores es relevante para la toma de decisiones por parte de las autoridades sanitarias, relacionado con la pandemia del Covid-19. De este modo, numerosas investigaciones se han llevado a cabo en el mundo para comprender este fenómeno, sus variantes y principales afectaciones en la salud de la población. En el presente estudio, por lo tanto, se realizó un análisis de conglomerados con base en las variables tasa de mortalidad, tasa de morbilidad y tasa de letalidad, así como con la densidad geográfica cantonal, para el periodo de la variante beta en Costa Rica, correspondiente a los meses de febrero a junio de 2021, y se eligieron tres métodos para obtener los grupos: k-medias, k-medoides y métodos difusos; así como dos tipos de distancias: euclídea y de Manhattan. Asimismo, se utilizó la suma de cuadrados dentro de grupos y el índice de Dunn para validar la conformación de los grupos. Se identificó que el método y la distancia que formaban los conglomerados de cantones más compactos, con una menor variabilidad intragrupo, fue k-medoides y Manhattan, respectivamente, debido a que su mayor robustez ante valores extremos. De los grupos formados, el clúster 1 posee un impacto moderado de la pandemia durante la variante en cuestión, y el grupo 2 y 3, un impacto bajo y alto, respectivamente. Asimismo, el grupo 1 y 2 están conformados en mayor medida por cantones no pertenecientes al Gran Área Metropolitana, en contraposición al grupo 3. Este

análisis proporciona ideas valiosas para las autoridades sanitarias al comprender los impactos de la pandemia de Covid-19 en las regiones costarricenses y contribuye al desarrollo de estrategias específicas para una gestión efectiva.

Introduction

Since the beginning of the Covid-19 pandemic, many efforts have been made to identify the main social, economic, and health repercussions of a pandemic. Likewise, numerous investigations have been directed to analyze the behavior of the virus and its respective variants. The question that arises is why it is relevant to continue studying this disease, after three years since its worldwide spread; the answer is based on the fact that the behavior of the virus is not so much related to seasonal effects [1], but rather to social dynamics and interventions or restrictions in health matters, and to the emergence of new variants and mutations [2].

Likewise, in the face of the increase in infections, health authorities have had to take measures aimed at mitigating the growing proliferation of the virus as much as possible. At this point, the importance of data knowledge and analysis as a tool to understand the behavior of the pandemic is highlighted, and based on this, decisions are made in favor of people's health. The disease surveillance constitutes the basis for the response to epidemics [3].

In addition to this, the analysis of the data to determine projections, trends, and indicators has been essential to identify advances and challenges in the management of the pandemic waves, related to the variants of the Covid-19. It is important to highlight that, due to the geospatial, socioeconomic, and access to health care characteristics of the population in different places, it is convenient to analyze the indicators associated with the measurement of an epidemic or pandemic in a more disaggregated way. For this reason, in the present document, the cantonal behavior of Costa Rica is analyzed according to the mortality, morbidity, and lethality rate, during the beta variant wave, which was shown to cause more severe disease and higher transmissibility than the original variant [2].

Therefore, the main objective of this study is to compare clustering methods and their respective distances, and to identify the one that best fits the data, based on the bioindicators already mentioned.

Methodology

The information involved in the analysis came from the epidemiological reports provided by the Costa Rican Ministry of Health (MINSA), which has been compiled by the Centro de Investigación Observatorio del Desarrollo (CIOdD) from the University of Costa Rica [4]. Therefore, data related to cantons, specifically daily deaths and cases were used as input for the calculation of relevant indicators such as the mortality, morbidity, and lethality rates. For their calculation, the projected population for 2021 of each canton was obtained from the Instituto Nacional de Estadística y Censos (INEC) of Costa Rica [5]. Moreover, the population density (obtained from INEC), was also considered as a variable for the clustering analysis, as a way of considering the average number of people who live in each canton, and who could eventually be infected.

Due to the behavior of the cases, cantons with average rates corresponding to extreme values are obtained, which, if ignored, generate the presence of new extreme values. In response to this, it is decided to explore three clustering methods to identify which of these forms more compact groups: (a) a method that calculates average points, (b) cantons in the central position (analogous to the median), and (c) a variant of the k-means method, in which each canton has a probability degree of belonging to a group. The three methods are: (1) K-means: it starts with a predetermined number of groups (k) and identifies their average (centroids). It assigns each observation to the cluster with the closest centroid, (2) K-medoids: it is similar to k-means, but instead of calculating the distance of each point to the group means, it identifies the observation in the central position of the cluster and assigns observations to that cluster, and (3) Fuzzy methods: it considers observations that can be associated with multiple clusters. Each observation has a degree of membership to each group [6].

It is important to highlight that two distances are selected to measure distances between cantons, which corresponds to the Euclidean and Manhattan ones, chosen for their ease of understanding, given that they have similar formula calculation. Also, Dunn index was used to validate the clusters, to quantitatively assess the quality of the data partition into groups. Moreover, to determine the number of clusters, the elbow method was used, that consists in selecting the number of groups in which the within-cluster sum of squared decreases. Also, the study period selected corresponds to the third pandemic wave related to the beta variant, which covered the dates between February 21, 2021, and July 28 of that same year; this period was chosen taken into consideration the characteristics of the beta variant, so the analysis could be used as a reference for future pandemic events under similar conditions.

Furthermore, the analyses were carried out using the R software [7] and the Kabré supercomputer.

Results and discussion

Based on the elbow method, the number of clusters suggested are three. From Table 1, it is possible to identify the values of the sum of squares within cluster for each canton division method and each distance.

Table 1. Sum of squares within cluster by method and distance.

Method	Distance (in millons)	
	Euclidean	Manhattan
k-means	193	193
k-medoids	197	180
Fuzzy methods	193	189

From table 1, it is determined that the minimum sum of square within cluster is 180 million, corresponding to k-medoids method with the Manhattan distance. This means that this method and distance produces clusters with cantons that are more similar to each other, related to the division variables. Moreover, for validating the clusters, the dunn index is shown in Table 2:

Table 2. Dunn index values by method and distance.

Method	Dunn Index	
	Euclidean	Manhattan
k-means	0.066	0.054
k-medoids	0.068	0.072
Fuzzy methods	0.066	0.044

It is possible to identify from Table 2 that the highest Dunn index is 0.072 corresponding to k-medoids with Manhattan distance. In this case, a higher Dunn index means that the groups have a lower intracluster variability. Therefore, both, the sum of squares within clusters and the Dunn index show that k-medoids with the Manhattan distance is the one that produces the most accurate results; this is because it is more robust to extreme values than the method of k-means. The same behavior is observed in an analogous way with respect to the fuzzy method because this is a variation of the k-means method. However, it is observed that the Dunn index is still small, which indicates that, although despite being the method that generates the greatest similarity between cantons, there is still a lot of dispersion within the groups.

Descriptively, the clustering analysis using the k-medoids method and Manhattan distance reveals distinct patterns among clusters. Cluster 1, comprising mostly non-Greater Metropolitan Area (GAM) cantons, particularly from Alajuela, Guanacaste, and Puntarenas, exhibits a high morbidity rate and moderate mortality and lethality rates, suggesting a moderate impact of the pandemic during the beta variant. Similarly, cluster 2, dominated by non-GAM cantons, especially from Cartago and Limón, shows a moderate morbidity and lethality rate, and a low mortality rate, implying a lower impact of the pandemic. In contrast, cluster 3, primarily composed of GAM cantons, particularly from San José and Heredia, demonstrates high mortality, moderate morbidity, and high lethality rates, indicating an impact with a higher ratio of deaths to infections compared to other clusters.

Conclusions and/or recommendations

It is concluded that the method that performs the best separation of groups is the k-medoids with the Manhattan distance, given that it has the lowest sum of squares and the highest Dunn index. These results coincide with those found by various authors in previous studies [8]. However, it is identified that there is still a high variability between the cantons within the groups, which is due to the presence of extreme values, which occur regularly in a bioinfectious events such as a pandemic. It is suggested to carry out an analysis with other variables, in order to identify the main characteristics of the clusters in terms of health authorities being able to identify the behavior and characterization of the clusters formed.

Acknowledgments

Thanks to the CIOdD for compiling COVID-19 data in Costa Rica, and to its researchers for their assistance in a preliminary analysis for this article.

References

[1] G. L. Vasconcelos, A. A. Brum, F. A. Almeida, A. M. Macêdo, G. C. Duarte-Filho y R. Ospina, «Standard and Anomalous Waves of COVID-19: A Multiple-Wave Growth Model for Epidemics,» Brazilian Journal of Physics, pp. 1867-1883, 2022.

[2] J. L. Jacobs, G. Haidar y J. W. W. Mellors, «COVID-19: Challenges of Viral Variants,» Annual Review of Medicine, vol. 74, pp. 31-53, 2023.

[3] N. Pearce, J. P. Vandenbroucke, T. J. VanderWeele y S. Greenland, «Accurate Statistics on COVID-19 Are Essential for Policy Guidance and Decisions,» AJPH, vol. 110, n° 7, pp. 949-951, 2020.

[4] Universidad de Costa Rica. Centro de Investigación Observatorio del Desarrollo, «Costa Rica,» 30 05 2022. [En línea]. Available: https://app.powerbi.com/view?r=eyJrIjoiMjU3M2NkNjQtMGIyOS00ZjRmLWE3NjY-tNDE2OWNkZjIxZTdjIiwidCI6ImFkNjNmZDZmLWE4OTctNDljZS1hZWU5LTRmYzYxNzY1NjY4YSJ9&pageName=ReportSection. [Último acceso: 10 09 2023].

[5] Instituto Nacional de Estadística y Censos, Costa Rica «Estadísticas demográficas. 2011 – 2025. Proyecciones nacionales. Población total proyectada al 30 de junio por grupos de edades, según provincia, cantón, distrito y sexo,» [En línea].

[6] F. Klawonn, R. Kruse y R. Winkler, «Fuzzy Clustering: More than just fuzzification,» Fuzzy Sets and Systems, vol. 281, pp. 272-279, 2015.

[7] R. C. Team, R: A language and environment for statistical computing, Vienna, Austria: R Foundation for Statistical Computing.

[8] Suwanda, R., Syahputra, Z., y Zamzami, E. M. «Analysis of euclidean distance and manhattan distance in the K-means algorithm for variations number of centroid», In Journal of Physics: Conference Series, vol 1566, 2020.

Preliminary analysis of acoustic detection of the Red-throated Caracara in northern Costa Rica

Análisis preliminar de detección acústica del Caracara Avispero en el norte de Costa Rica

Roberto Vargas-Masís[1], Diego Quesada[2]

Vargas-Masís, R; Quesada, D. Preliminary analysis of acoustic detection of the red-throated Caracara in northern Costa Rica. *Tecnología en Marcha*. Vol. 37, special issue. August, 2024. IEEE International Conference on BioInspired Processing. Pag. 81-86.

https://doi.org/10.18845/tm.v37i7.7303

[1] Laboratorio de Investigación e Innovación Tecnológica, Universidad Estatal a Distancia. Costa Rica.
✉ rovargas@uned.ac.cr
iD https://orcid.org/0000-0003-1244-4381

[2] Birding Experiences. Caracara Project. Heredia, Costa Rica.
✉ diego@caracaracr.com
iD https://orcid.org/009-009-6988-5621

Keywords

Birds; acoustic detection; machine learning; passive acoustic monitoring; sound.

Abstract

The noisy Red-Throated Caracara (*Ibycter americanus*) is a species whose population has inexplicably declined across much of its range and is now rare in the Pacific and Caribbean slopes of Costa Rica. Advances in automatic acoustic detection have transformed bird ecology, allowing researchers to analyze bird populations using pattern matching algorithms, machine learning, and random forest models. Although these studies are limited in the country, it represents an area with great interdisciplinary potential for technological advances. This study focused on the use of Pattern Matching to detect the presence of the Red-Throated Caracara in northern Costa Rica using a large number of sound recordings and its validation with metrics such as Accuracy, Precision Negative predictive value, Sensitivity, Specificity and Unweighted average recall. The results showed a moderate performance of the model by obtaining accuracy and precision values of 0.71 compared to the values obtained in other investigations in which the reported model was used. Therefore, we suggest exploring new techniques and methods to improve the detection of the species, considering the particular acoustic structure, the repertoire of sounds of the species and similarities with vocalizations of other species. This similarity could indicate a supposed anti-predator defense behavior by "imitating" the sounds of other species with which it shares habitat. To optimize this acoustic detection, we recommend using complementary techniques such as noise filters that improve the quality and precision of the data.

Palabras clave

Aves; detección acústica; aprendizaje automático; monitoreo acústico pasivo; sonido.

Resumen

El ruidoso Caracara de Garganta Roja (*Ibycter americanus*) es una especie cuya población ha disminuido inexplicablemente a lo largo de su área de distribución y es ahora rara en ambas vertientes de Costa Rica. Los avances en la detección acústica automática han transformado la ecología de las aves, permitiendo a los investigadores analizar poblaciones de aves utilizando algoritmos de coincidencia de patrones, aprendizaje automático y modelos forestales aleatorios. Aunque estos estudios son limitados en el país, representan un área con gran potencial interdisciplinario para avances tecnológicos. Este estudio se centró en el uso de algoritmos de coincidencia de patrones para detectar la presencia del Caracara avispero en el norte de Costa Rica utilizando una gran cantidad de grabaciones sonoras y su validación con métricas como Exactitud, Precisión, Valor predictivo negativo, Sensibilidad, Especificidad y Recall medio no ponderado. Los resultados mostraron un desempeño moderado del modelo al obtener valores de exactitud y precisión de 0.71 en comparación con los valores obtenidos en otras investigaciones en las que se utilizó el modelo reportado. Por ello sugerimos explorar nuevas técnicas y métodos para mejorar la detección de la especie, considerando la particular estructura acústica, el repertorio de sonidos de la especie y similitudes con vocalizaciones de otras especies. Esta similitud podría indicar un supuesto comportamiento de defensa anti-depredador al "imitar" los sonidos de otras especies con las que comparte el hábitat. Para optimizar esta detección acústica recomendamos hacer uso de técnicas complementarias como los filtros de ruido que mejoren la calidad y precisión de los datos.

Introduction

The lowlands of the Caribbean slope of northern Costa Rica constitute one of the priority hotspots for biodiversity conservation in Mesoamerica. Nevertheless, the landscape has undergone a process of strong fragmentation that threatens its connectivity between protected areas in Costa Rica and southeastern Nicaragua [1] which impacts species with particular conservation status in these habitats [2].

Red-Throated Caracara (*Ibycter americanus*) is a noisy and widespread forest caracara. Pairs and family groups are very vocal, and their calls often can be heard from great distances in lowlands and lower mountain areas from southern Mexico to southern Brazil [3].

The species is a highly territorial cooperative breeder and groups often stay together for many years. In Costa Rica, formerly widespread and common in the forests but its population has inexplicably disappeared from most of this range, although in recent years its population apparently increase again in northeastern Costa Rica and Osa Peninsula [4].

Important advances in the field of bird ecology have been made possible by techniques such as automatic acoustic detection to understand many aspects of bird populations and their ecosystems through sound. These techniques not only improve scientific understanding of bird communities, but also hold great promise for biodiversity conservation [5].

Methods such as Pattern Matching Algorithms, Machine Learning, Random Forest Models (RFM), Hidden Markov Models, Support Vector Machines, Mel-Frequency Cepstral Coefficients, and Convolutional Recurrent Neural Networks are leading the field of automatic bird acoustic detection [7, 8].

In Costa Rica some studies of automated acoustic detection of birds have been developed, [9, 10]. In this study, we used a Pattern Matching method to label the presence of the Red-throated Caracara in northern of Costa Rica through a preliminary analysis of the acoustic detection for recommendations of future research and analysis.

Methodology

Data Collection

The study site is located in the Monterrey district of San Carlos, Costa Rica (10.642271° N, -84.664404° O). A farm of approximately 250 hectares used as a cattle ranch for beef. The area has irregular topography with pastures that are surrounded by fragmented patches of tropical lowland rainforest, with high precipitation, humidity and temperatures that average 30 degrees Celsius throughout the year.

The acoustic soundscape was monitored using four Wildlife Acoustic Micro recorders, placed in strategic sites where the species was previously observed. A total of 221,016 recordings of one minute were uploaded in ARBIMON platform.

Data Analysis

Then we used a sample of 64,736 recordings and the pattern of one vocalization of the caracara to run a Pattern Matching feature with 0.1 threshold value and one match per recording to assist the validation process. The sound selected to this study presents a Bandwidth of 6533 Hz, and its length is about 0.53 s (Figure 1).

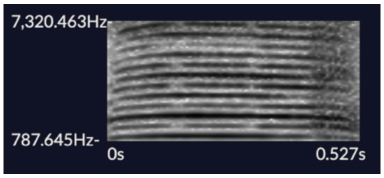


Figure 1. Training set pattern of the sound of Red-throated Caracara.

We used the output of the Pattern Matching to check and mark the presence (30 tags) and absence (40 tags) of the Caracara. We used 70/30 ratio for training and validation data. The presence or absence in the validations are contrasted with the ROI used in the Pattern Matching function (template of the sound of interest) to validate the detections in the platform (True Positive or *TP*, True Negative or *TN*, False Positive *FP* and False Negative *FN*).

To perform model validation, we used a confusion matrix that describes the performance of the binary classifier (presence or absence of the species). Taking as reference the research of Vargas-Masís and collaborators [9, 10] we calculated the same metrics to validate the model (the formula for each metric is described in reference papers) using Accuracy (*Ac*), Precision (*Pr*), Negative predictive value (*Npv*), Sensitivity or recall (*Se*), Specificity (*Sp*), F_1Score measure as 2*(precision*recall)/(precision+recall), Area Under the Curve (AUC) and Unweighted average recall (*UAR*).

Results

Model performance

The confusion matrix for the test dataset based on validations of the RFM shows for that TP (5) was higher than FP (2) and FN (4) was lower than TN (10). This means that the classifier correctly predicts the presence in a higher number of true positives and true negative records compared with the false positives and false negatives in the evaluated models.

Table 1. *Accuracy (Ac), Precision (Pr), Negative predictive value (Npv), Sensitivity (Se), Specificity (Sp), F_1Score Area Under the Curve (AUC) and Unweighted average recall (UAR) of Red-Throated Caracara in northern of Costa Rica.*

Metrics	Ac	Pr	Npv	Se	Sp	F_1Score	AUC	AUR
Caracara	0.71	0.71	0.71	0.56	0.83	0.63	0.87	0.69

Table 1 shows the results of the confusion matrix that will subsequently allow measuring the performance of the RFM. For the proposed model, the ratio of TP and FP, as well as FN and TN agree with a classifier that will correctly predict the presence in a larger number of true positives and true negatives of a larger dataset considering the selected acoustic pattern.

In the case of this model (Table I), the *Ac* was lower than the *Pr* and in this case it will correctly predict about 71% of the data for the species. An *Npv* of 71% indicates that the test has a moderate ability to rule out the identification of the acoustic pattern in negative cases. It is not a high *Npv*, however, it is still useful for acoustic pattern identification.

In addition, *Se* shows low ability of the model to find positive cases with only 56% of True Positives. For the *Sp* indicator it shows a high ability of the model to predict true negatives in 83% of the cases. *F1score* combines accuracy and sensitivity into a single value, and the result (63%) shows moderate model performance.

Finally, *AUC* presents a value between 0.80-0.90 which is considered good for discriminating in the model between positive and negative classes. But in the case of *UAR*, it was obtained that 69% of the classification model performs moderately well in terms of its ability to recognize the acoustic pattern of the species in a dataset with unbalanced classes.

Conclusions

In general, a moderate performance was obtained in terms of its ability to recognize the acoustic pattern of the species in the analyzed data set, it is important to study the technical factors of this vocalization that could affect its detection.

This sound is composed of a large number of harmonics that in addition to its intensity (dB) and frequency bandwidth (Hz) resembles to a great extent the sounds of other common species of the Psittacidae family (macaws, parrots, and parakeets). This similarity could indicate a hypothesized anti-predator defense behavior by “mimicking” the sounds of predators or prey in order to go unnoticed in the environment [11].

Based on the performance indicators of the model, we recommend studying other techniques that may benefit the detection of Caracara. Mel-Frequency Cepstral Coefficients (MFCCs) has presents benefits in detecting complex sounds such as the human voice and bird sounds [12] with a large number of harmonics.

Although when compared to recent studies, differences in model performance metrics are found, studies such as Sun and collaborator show similar values that can be improved by increasing the amount of data available using a convolutional neural network model [13].

In future projects it is important to consider a combination of techniques or through an ensemble approach [8] that integrate the entire repertoire of sounds of *I. americanus* to better characterize the acoustic activity of the species [14]. It is even necessary to consider the use of different detection techniques to take into consideration the acoustic structure which, although it makes the work more complex, could benefit its detection.

For bioacoustic data we recommend using typical ratios (70-30 or 80-20 in training-validation) depending on the amount of data available because this provides sufficient data to effectively train the model, validate its performance, and assess its generalizability on unvalidated data [15]. It is necessary to apply noise filtering techniques that can improve the detection process in recordings that have a greater number of environmental interactions (wind, rain, insects) that can interfere in the correct detection [16].

Through this acoustic monitoring it has been possible to determine the use that these individuals make of the site during this monitoring, which is key to better understand their behaviors, habitat preferences and resource needs, interaction with other species and thus better understand the dynamics among these populations.

The challenge of this project is to determine where and when *I. americanus* moves to and from given the area’s need for resources. Obtaining more data on its distribution, time spent in the area and vocalization periods will help to create proposals for habitat management of the forest matrix that this species needs for its activities.

Acknowledgements

We thank Mr. Marconi Cedeño and Mr. Arnoldo Cedeño for permissions and all the support to develop the whole research process. A special thanks to Antonio “Toño” Dávila because all his sightings and alerts about the species were key to this project. To Sean McCann, Jehudy

Carballo, Mario Salazar, Pablo Camacho, Anthony Arce, Gerald Pereira, Priscila Salazar, Jorge Campos, Saúl Herrera and David Quesada for all the help in field work and suggestions. To Rainforest connection for the access, backup, and analysis in Arbimon platform.

References

[1] Ramírez-Albores, Jorge E., et al. "Insights for protection of high species richness areas for the conservation of Mesoamerican endemic birds". *Diversity and Distributions*, vol. 27, no 1, p. 18-33,2021.

[2] Chassot, Olivier; Monge-Arias, G. "Connectivity conservation of the Great Green Macaw's landscape in Costa Rica and Nicaragua (1994–2012)". Parks, vol. 18, no 1, p. 61-69, 2012.

[3] Mccann, Sean Michael. "The bird who kicked the wasp's nest: Red-throated Caracara predation, nesting and territorial behavior". Doctoral Thesis. Simon Fraser University, 2014.

[4] Dyer, Dale; Howell, Steve NG. "Birds of Costa Rica". Princeton University Press, vol. 1, no 1, 2023.

[5] Bardeli, Rolf, et al. "Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring". *Pattern Recognition Letters*, vol. 31, no 12, p. 1524-1534, 2010.

[6] Shonfield, Julia; Bayne, Erin M. "Autonomous recording units in avian ecological research: current use and future applications". *Avian Conservation & Ecology*, vol. 12, no 1, 2017.

[7] XIE, Jiangjian, et al. "A review of automatic recognition technology for bird vocalizations in the deep learning era". *Ecological Informatics*, p. 101927, 2022.

[8] Castro, Jorge; Vargas-Masís, Roberto; Alfaro-Rojas, Danny. "Deep multiple instance learning ensemble for the acoustic detection of tropical birds". *19th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE.* p. 264-269, 2020.

[9] Vargas-Masís, Roberto, et al. "Acoustic detection of Red-capped Manakin (*Ceratopipra mentalis*) in Sarapiquí, Costa Rica". *3rd International Conference on BioInspired Processing (BIP). IEEE*. p. 1-5, 2021.

[10] Vargas-Masís, Roberto, et al. "Automated bird acoustic detection at Las Arrieras Nature Reserve in Sarapiquí, Costa Rica". *4th International Conference on BioInspired Processing (BIP). IEEE.* p. 1-8, 2022.

[11] Thiollay, Jean-Marc; Jullien, Mathilde. "Flocking behaviour of foraging birds in a neotropical rain forest and the antipredator defence hypothesis". *Ibis*, vol. 140, no 3, p. 382-394, 1998.

[12] Ramirez, Angel David Pedroza, et al. "A comparative between mel frequency cepstral coefficients (MFCC) and inverse mel frequency cepstral coefficients (IMFCC) features for an automatic bird species recognition system". *Latin American Conference on Computational Intelligence (LA-CCI). IEEE*, 2018. p. 1-4, 2018.

[13] Sun, Yuren, et al. "Classification of animal sounds in a hyperdiverse rainforest using convolutional neural networks with data augmentation." *Ecological Indicators* 145 (2022): 109621.

[14] Brooker, Stuart A., et al. "Automated detection and classification of birdsong: An ensemble approach". *Ecological Indicators*, vol. 117, p. 106609, 2020.

[15] Muraina, Ismail. "Ideal dataset splitting ratios in machine learning algorithms: general concerns for data scientists and data analysts". *7th International Mardin Artuklu Scientific Research Conference*, 2022.

[16] Bardeli, Rolf, et al. "Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring". *Pattern Recognition Letters*, vol. 31, no 12, p. 1524-1534, 2010.

Artificial intelligence, machine learning and GIS in environmental engineering: current trends

Inteligencia artificial, machine learning y SIG en ingeniería ambiental: tendencias actuales

Laura Hernández-Alpízar[1], José Andrés Gómez-Mejía[2], María Belén Argüello-Vega[3]

Hernández-Alpízar, L; Gómez-Mejía, J. A; Argüello-Vega, M. B. Artificial intelligence, machine learning and GIS in environmental engineering: current trends. *Tecnología en Marcha*. Vol. 37, special issue. August, 2024. IEEE International Conference on BioInspired Processing. Pag. 87-96.

https://doi.org/10.18845/tm.v37i7.7304

[1] Instituto Tecnológico de Costa Rica. Costa Rica.
lahernandez@itcr.ac.cr
https://orcid.org/0000-0002-9193-8429

[2] Instituto Tecnológico de Costa Rica. Costa Rica.
jagomezm@estudiantec.cr
https://orcid.org/0009-0005-1769-7283

[3] Instituto Tecnológico de Costa Rica. Costa Rica.
belenarguello@estudiantec.cr
https://orcid.org/0009-0006-1224-2658

Keywords

Computational tools; database; water; energy; air; solutions.

Abstract

Recent advances in Artificial Intelligence (AI), Machine Learning (ML), and Geographic Information Systems (GIS) have significantly enhanced our understanding of environmental issues. This review analyzes publications from the IEEE Xplore Digital Library to assess the growing expertise in these fields. By applying filters based on year, technique, and keywords such as water, air, soil, climate change, energy, and waste, we visualize the evolving application of these technologies across key environmental topics. Our findings offer scientific guidance on the most relevant applications and highlight areas in need of further investigation. A detailed review of the literature also reveals the connection between different domains and their impact. This work intents to promote ongoing research and serve as a critical resource in the search for solutions to environmental challenges

Palabras clave

Herramientas computacionales; base de datos; agua; energía; aire; soluciones.

Resumen

Los avances recientes en inteligencia artificial (IA), aprendizaje automático (ML) y sistemas de información geográfica (SIG) han mejorado significativamente nuestra comprensión de los problemas ambientales. Esta revisión analiza las publicaciones de la Biblioteca Digital IEEE Xplore para evaluar la creciente experiencia en estos campos. Al aplicar filtros basados en año, técnica y palabras clave como agua, aire, suelo, cambio climático, energía y residuos, visualizamos la aplicación cambiante de estas tecnologías en temas ambientales clave. Nuestros hallazgos ofrecen orientación científica sobre las aplicaciones más relevantes y resaltan áreas que necesitan más investigación. Una revisión detallada de la literatura también revela la conexión entre diferentes dominios y su impacto. Este trabajo tiene como objetivo fomentar la investigación en curso y servir como un recurso crítico en la búsqueda de soluciones a los desafíos ambientales.

Introduction

There are three large groups of computational tools constantly mentioned in the field of environmental engineering for the management of scientific data in their areas of interest, these are: Artificial Intelligence (AI) [1], Machine Learning (ML) [2] and Geographic Information Systems (GIS) [3]. These tools facilitate data analysis and pattern recognition, facilitating cost-effective decision-making compared to traditional sampling and laboratory methods. They also aid in understanding large, interconnected territories and complex matrices, allowing for the early detection of environmental issues [4]-[7].

AI, ML and GIS have been used to address complex issues in areas of environmental engineering such as water quality [8]-[9], energy management [10], air and soil pollution [11]-[12], waste management [13] and climate change [14]-[15]. This work aims to promote the development of collaborative work solutions in the related areas of environmental and computational engineering by visualizing potential areas of their development, through a systematic literature review of publications in *IEEEXplore* Digital Library.

Methodology

The *IEEE Xplore* database was chosen because it makes interrelations between various engineering fields and is also a recognized editorial in technology publications. Likewise, enables to search not only with keywords, but with publication topics. A search was performed using keywords and selected publication topics, from 2012 to 2022 as an arbitrary time range, for three major groups of data treatment tools: AI, ML and GIS. For each of these tools, a search was carried out by areas of interest in environmental engineering according to [16]. Those areas are water, air, soil, climate change, energy and waste. The number of existing publications in the database was quantified for each year. The keywords used were: "tool" (e.g., "machine learning"), "environmental engineering" and "area" (e.g., "water"). The search was refined with the publication topics according to table 1.

Table 1. Publication topics used as filter in IEEE Xplore Digital Library for each area and tool.

Area	ML	AI	GIS
Water	learning (artificial intelligence), environmental science computing, water quality, regression analysis, remote sensing, water resources, Internet of Things, rivers, water supply, water pollution, environmental monitoring (geophysics), deep learning (artificial intelligence), lakes	learning (artificial intelligence), environmental science computing, water quality, neural nets, regression analysis, water resources, Internet of Things, deep learning (artificial intelligence), artificial intelligence, image classification, hydrological techniques, water supply, convolutional neural nets, feature extraction, water pollution, time series, rivers, wastewater treatment	geographic information systems, remote sensing, environmental science computing, water resources, water quality, rivers, hydrological techniques, water pollution, geophysical image processing, terrain mapping, groundwater, rain, geophysics computing, lakes, water supply, environmental monitoring (geophysics)
Air	learning (artificial intelligence), air pollution, environmental science computing, air quality, regression analysis, neural nets, Internet of Things, air pollution control, data analysis, deep learning (artificial intelligence), environmental monitoring (geophysics), air pollution measurement	learning (artificial intelligence), environmental science computing, air pollution, air quality, neural nets, deep learning (artificial intelligence), regression analysis, air pollution control, recurrent neural nets, Internet of Things, data analysis, artificial intelligence, air pollution.	geographic information systems, air pollution, environmental management, environmental science computing, remote sensing, Gaussian distribution, Gaussian processes, coal, contamination, environmental degradation, environmental factors, geophysical techniques
Soil	learning (artificial intelligence), soil, regression analysis, Internet of Things, fertilizers, random forests, remote sensing, environmental science computing, deep learning (artificial intelligence), agricultural products, mean square error methods	agriculture, artificial intelligence, crops, irrigation, soil, agrochemicals, deep learning (artificial intelligence), farming, fertilizers, learning (artificial intelligence), neural nets, pesticides, plant diseases, regression analysis	geographic information systems, remote sensing, land use planning, soil, terrain mapping, Global Positioning System, agricultural products, environmental degradation, erosion, forestry, geophysical image processing, geophysical techniques, geophysics computing, irrigation, planning

Area	ML	AI	GIS
Climate Change	learning (artificial intelligence), environmental science computing, neural nets, Internet of Things, regression analysis, climate mitigation, geophysical image processing, air pollution, global warming, remote sensing, data analysis, ecology, rivers	learning (artificial intelligence), neural nets, climate mitigation, deep learning (artificial intelligence), Internet of Things, artificial intelligence, remote sensing, ecology, global warming, regression analysis, convolutional neural nets, image classification, recurrent neural nets, support vector machines, environmental factors	geophysical information systems, remote sensing, climatology, terrain mapping, vegetation mapping, ecology, environmental science computing, geophysical image processing, vegetation, atmospheric precipitation, atmospheric temperature, climate mitigation, environmental management, floods, disasters, global warming
Energy	learning (artificial intelligence), energy consumption, optimization, environmental science computing, energy management systems	learning (artificial intelligence), deep learning (artificial intelligence), energy consumption, energy conservation, photovoltaic power systems, regression analysis, power grids, distributed power generation, renewable energy sources, energy management systems, load forecasting, Internet of Things, artificial intelligence	geographic information systems, bioenergy conversion, energy management systems, power engineering computing, power generation planning, power system management, power system planning, renewable energy sources, smart power grids
Waste	learning (artificial intelligence), environmental science computing, recycling, Internet of Things, waste management, refuse disposal, deep learning (artificial intelligence), image classification, waste disposal, regression analysis, artificial intelligence, waste recovery, waste reduction	learning (artificial intelligence), recycling, deep learning (artificial intelligence), waste management, Internet of Things, waste disposal, artificial intelligence, object detection, waste recovery, waste handling, feature extraction, waste reduction, industrial waste, municipal solid waste	geographic information systems, remote sensing, coal, environmental factors

Results and discussion

The analysis reveals an exponential increase in AI applications in environmental engineering over the past decade, particularly from 2018 (Figure 1). The primary application area is energy (40%), with notable trends in forecasting models, such as those for wind turbine energy production [17], energy consumption [18], and renewable energy generation [19]. Important studies on IA applications in energy management systems that propose environmental solutions are also found [20]-[21].

The second area of high IA publications in environmental engineering is air, representing 21% of the total. These applications include proposals based on pollution and quality data monitoring [22]-[23]. Additionally, water ranked third in the literature search, with publications such as monitoring and predictive models of water quality [24]-[25].

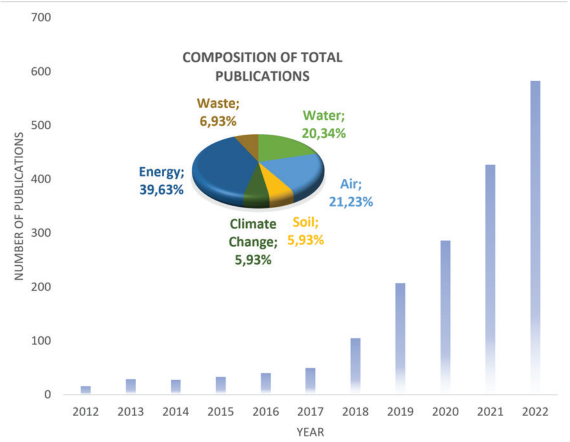


Figure 1. Number of publications of AI applications in environmental engineering found in IEEE Xplore Digital Library (access date: 08/26/2023). The inset shows the areas' composition of the total number of articles.

Figure 2 depicts the exponential growth in Machine Learning applications in environmental engineering, with a noticeable acceleration since 2018. These publications predominantly focus on energy (33%), air quality (25%), and water quality (19%).

In the energy sector, ML has been applied to enhance energy management and predict consumption for homes, buildings, and cities [26]-[32], including renewable energy sources [33]-[35]. For air quality, ML is applied to predict air pollution [36]-[40], often integrated with the Internet of Things (IoT) [41]-[43]. Other studies explore the link between air pollutants and diseases using ML [44]-[45]. Technical aspects of ML algorithms for air pollution prediction have been reviewed [46]. Water quality applications involve the use of IoT and ML for real-time monitoring and prediction [47]-[51]. Algorithms for river quality assessment have been developed [52]-[53], and ML is applied to analyze drinking water quality [54]-[57]. Gai and Yang [58] provide a comprehensive review of ML-based water quality prediction methods and discuss future trends.

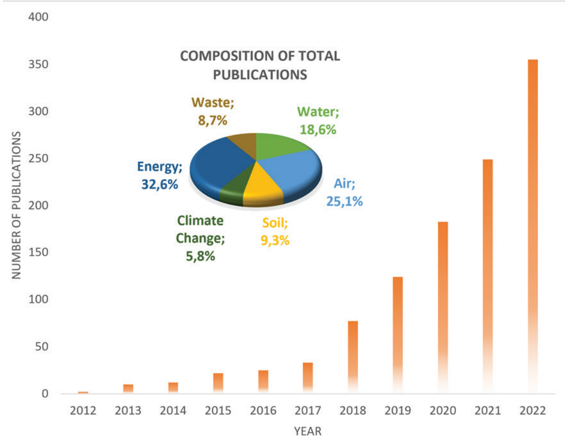


Figure 2. Number of publications of ML applications in environmental engineering found in IEEE Xplore Digital Library (access date: 08/26/2023). The inset shows the areas' composition of the total number of articles.

Regarding the trend in the number of publications in the database on applications of Geographic Information Systems (figure 3), this did not present an exponential growth as in the case of AI and ML. There was only a significant increase between the years 2018-2021. The water area had the most publications (54%), with research on monitoring of water quality and prediction systems [59]-[60]. The energy area was the second highest incidence (15%), where GIS have been used largely for the selection of optimal locations for renewable energy systems [61]-[62].

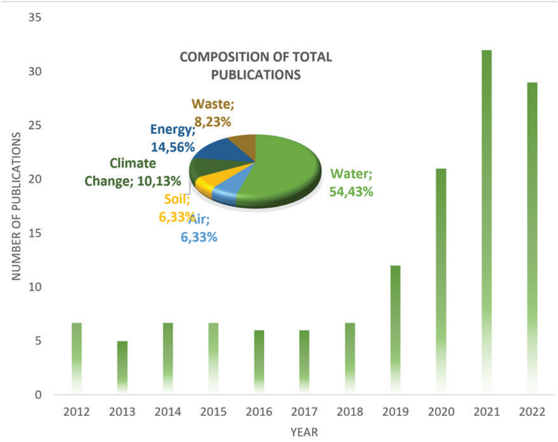


Figure 3. Number of publications of GIS applications in environmental engineering found in IEEE Xplore Digital Library (access date: 08/26/2023). The inset shows the areas' composition of the total number of articles.

Conclusions and recommendations

In this limited but detailed review experiment by topics, tool and area of knowledge, it is shown how the management and modeling of environmental information has intensified in the last decade, heading towards the main environmental and sustainability problems of the Earth. Although it is recommended to investigate other databases to analyze trends that are not so present in IEEE and extend the research to other computational tools, the results are provided to the scientific community to strengthen investigative work initiatives within the development framework of environmental and computational engineering and global concerns.

Acknowledgments

The authors would like to thank the Research Center for Environmental Protection (CIPA) of the Technological Institute of Costa Rica for financial and laboratory support.

References

[1] E. K. Nti *et al*, "Environmental sustainability technologies in biodiversity, energy, transportation and water management using artificial intelligence: A systematic review," *Sustainable Futures,* vol. 4, pp. 100068, 2022. DOI: 10.1016/j.sftr.2022.100068.

[2] S. Zhong *et al*, "Machine Learning: New Ideas and Tools in Environmental Science and Engineering," *Environ. Sci. Technol.,* vol. 55, *(19),* pp. 12741-12754, 2021. DOI: 10.1021/acs.est.1c01339.

[3] M. M. Nowak *et al*, "Mobile GIS applications for environmental field surveys: A state of the art," *Global Ecology and Conservation,* vol. 23, pp. e01089, 2020. DOI: 10.1016/j.gecco.2020.e01089.

[4] P. Tahmasebi *et al*, "Machine learning in geo- and environmental sciences: From small to large scale," *Adv. Water Resour.,* vol. 142, pp. 103619, 2020. DOI: 10.1016/j.advwatres.2020.103619.

[5] S. S. Gill *et al*, "AI for next generation computing: Emerging trends and future directions," *Internet of Things,* vol. 19, pp. 100514, 2022. DOI: 10.1016/j.iot.2022.100514.

[6] S. Gupta *et al*, "Data Analytics for Environmental Science and Engineering Research," *Environ. Sci. Technol.,* vol. 55, *(16),* pp. 10895-10907, 2021. DOI: 10.1021/acs.est.1c01026.

[7] G. Lü *et al*, "Reflections and speculations on the progress in Geographic Information Systems (GIS): a geographic perspective," *Int. J. Geogr. Inf. Sci.,* vol. 33, *(2),* pp. 346-367, 2019. DOI: 10.1080/13658816.2018.1533136.

[8] M. Zhu *et al*, "A review of the application of machine learning in water quality evaluation," *Eco-Environment & Health,* vol. 1, *(2),* pp. 107-116, 2022. DOI: 10.1016/j.eehl.2022.06.001.

[9] Z. Liu *et al*, "Remote sensing and geostatistics in urban water-resource monitoring: a review," *Mar. Freshwater Res.,* vol. 74, *(10),* pp. 747-765, 2023. DOI: 10.1071/MF22167.

[10] T. Ahmad *et al*, "Artificial intelligence in sustainable energy industry: Status Quo, challenges and opportunities," *J. Clean. Prod.,* vol. 289, pp. 125834, 2021. DOI: 10.1016/j.jclepro.2021.125834.

[11] A. Samad *et al*, "Air pollution prediction using machine learning techniques – An approach to replace existing monitoring stations with virtual monitoring stations," *Atmos. Environ.,* vol. 310, pp. 119987, 2023. DOI: 10.1016/j.atmosenv.2023.119987.

[12] S. Liu *et al*, "Status and environmental management of soil mercury pollution in China: A review," *J. Environ. Manage.,* vol. 277, pp. 111442, 2021. DOI: 10.1016/j.jenvman.2020.111442.

[13] M. Abdallah *et al*, "Artificial intelligence applications in solid waste management: A systematic research review," *Waste Manage.,* vol. 109, pp. 231-246, 2020. DOI: 10.1016/j.wasman.2020.04.057.

[14] L. Chen *et al*, "Artificial intelligence-based solutions for climate change: a review," *Environmental Chemistry Letters,* vol. 21, *(5),* pp. 2525-2557, 2023. DOI: 10.1007/s10311-023-01617-y.

[15] A. Balogun *et al*, "A review of the inter-correlation of climate change, air pollution and urban sustainability using novel machine learning algorithms and spatial information science," *Urban Climate,* vol. 40, pp. 100989, 2021. DOI: 10.1016/j.uclim.2021.100989.

[16] A. Alfaro-Barquero and S. Chinchilla-Brenes. "Preferencias y habilidades vocacionales de las Ingenierías Ambiental, Forestal y Seguridad Laboral e Higiene Ambiental: Vocational preferences and skills in Environmental Engineering, Forestry and Occupational Safety and Environmental Hygiene," *Revista Digital: Matemática*, Educación E Internet, vol. 20, (2), 2020. Available: https://tecdigital.tec.ac.cr/servicios/revistamatematica/ArticulosRevistaDigitalV20_n2_2020/RevistaDigital_AlfaroBrenes_V20_n2_2020/RevistaDigital_AlfaroBrenes_V20_n2_2020.pdf

[17] T. Bhardwaj, S. Mehenge and B. S. Revathi, "Wind Turbine Power Output Forecasting Using Artificial Intelligence," 2022 International Virtual Conference on Power Engineering Computing and Control: Developments in Electric Vehicles and Energy Sector for Sustainable Future (PECCON), Chennai, India, 2022, pp. 1-5, doi: 10.1109/PECCON55017.2022.9851008.

[18] T. C. Brito and M. A. Brito, "Forecasting of Energy Consumption : Artificial Intelligence Methods," 2022 17th Iberian Conference on Information Systems and Technologies (CISTI), Madrid, Spain, 2022, pp. 1-4, doi: 10.23919/CISTI54924.2022.9820078.

[19] Purwanto, Hermawan, Suherman, D. A. Widodo and N. Iksan, "Renewable Energy Generation Forecasting on Smart Home Micro Grid using Deep Neural Network," 2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS), Bandung, Indonesia, 2021, pp. 1-4, doi: 10.1109/AIMS52415.2021.9466089.

[20] Q. Sun, D. Wang, D. Ma and B. Huang, "Multi-objective energy management for we-energy in Energy Internet using reinforcement learning," 2017 IEEE Symposium Series on Computational Intelligence (SSCI), Honolulu, HI, USA, 2017, pp. 1-6, doi: 10.1109/SSCI.2017.8285243.

[21] A. A. Allal, K. Mansouri, M. Youssfi and M. Qbadou, "Toward a review of innovative solutions in the ship design and performance management for energy-saving and environmental protection," 2018 19th IEEE Mediterranean Electrotechnical Conference (MELECON), Marrakech, Morocco, 2018, pp. 115-118, doi: 10.1109/MELCON.2018.8379078.

[22] A. Srivastava, A. Ahmad, S. Kumar and M. A. Ahmad, "Air Pollution Data and Forecasting Data Monitored through Google Cloud Services by using Artificial Intelligence and Machine Learning," 2022 6th International Conference on Electronics, Communication and Aerospace Technology, Coimbatore, India, 2022, pp. 804-808, doi: 10.1109/ICECA55336.2022.10009293.

[23] A. Vishnubhatla, "IoT based Air Pollution Monitoring through Telit Bravo Kit," 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 2022, pp. 1751-1755, doi: 10.1109/ICAAIC53929.2022.9793252.

[24] J. Lan, P. Zhang and Y. Huang, "Application Research of Computer Artificial Intelligence Monitoring System in Surface Water Quality Measurement of Water Conservancy Industry," 2022 International Conference on Education, Network and Information Technology (ICENIT), Liverpool, United Kingdom, 2022, pp. 311-314, doi: 10.1109/ICENIT57306.2022.00075.

[25] N. Desai and Dhinesh Babu L.D, "Software sensor for potable water quality through qualitative and quantitative analysis using artificial intelligence," 2015 IEEE Technological Innovation in ICT for Agriculture and Rural Development (TIAR), Chennai, India, 2015, pp. 208-213, doi: 10.1109/TIAR.2015.7358559.

[26] T. C. Brito and M. A. Brito, "Forecasting of energy consumption : Artificial intelligence methods," *2022 17th Iberian Conference on Information Systems and Technologies (CISTI)*, Madrid, Spain, 2022, pp. 1-4, doi: 10.23919/CISTI54924.2022.9820078.

[27] R. G. Rajasekaran, S. Manikandaraj and R. Kamaleshwar, "Implementation of machine le arning algorithm for predicting user behavior and smart energy management," *2017 International Conference on Data Management, Analytics and Innovation (ICDMAI)*, Pune, India, 2017, pp. 24-30, doi: 10.1109/ICDMAI.2017.8073480.

[28] A. Talwariya et al, "Domestic energy consumption forecasting using machine learning," *2022 7th International Conference on Smart and Sustainable Technologies (SpliTech),* 2022, . DOI: 10.23919/SpliTech55088.2022.9854296.

[29] F. Fakour et al, "Machine learning & uncertainty quantification: Application in building energy consumption," *2022 Annual Reliability and Maintainability Symposium (RAMS*), 2022, . DOI: 10.1109/RAMS51457.2022.9893988.

[30] X. Yang et al, "A forecasting method of air conditioning energy consumption based on extreme learning machine algorithm," *2017 6th Data Driven Control and Learning Systems (DDCLS),* 2017, . DOI: 10.1109/DDCLS.2017.8068050.

[31] Z. Wu and W. Chu, "Sampling strategy analysis of machine learning models for energy consumption prediction," *2021 IEEE 9th International Conference on Smart Energy Grid Engineering (SEGE),* 2021, . DOI: 10.1109/SEGE52446.2021.9534987.

[32] A. Prasad et al, "Analyzing land use change and climate data to forecast energy demand for a smart environment," 2021 9th International Renewable and Sustainable Energy Conference (IRSEC), Morocco, 2021, pp. 1-6, doi: 10.1109/IRSEC53969.2021.9741210.

[33] P. Rezaei et al, "A novel energy management scheme for a microgrid with renewable energy sources considering uncertainties and demand response," *2022 12th Smart Grid Conference (SGC)*, 2022, . DOI: 10.1109/SGC58052.2022.9998980.

[34] A. Alavi-Koosha et al, "Trend curve- and machine learning-based renewable energy development forecast," *2022 12th Smart Grid Conference (SGC)*, 2022, . DOI: 10.1109/SGC58052.2022.9998917.

[35] A. Shahab and M. P. Singh, "Comparative analysis of different machine learning algorithms in classification of suitability of renewable energy resource," *2019 International Conference on Communication and Signal Processing (ICCSP)*, 2019, . DOI: 10.1109/ICCSP.2019.8697969.

[36] B. D. Parameshachari et al, "Prediction and analysis of air quality index using machine learning algorithms," *2022 IEEE International Conference on Data Science and Information System (ICDSIS)*, 2022, . DOI: 10.1109/ICDSIS55133.2022.9915802.

[37] K. M. O. V. K. Kekulanadara, B. T. G. S. Kumara and B. Kuhaneswaran, "Comparative analysis of machine learning algorithms for predicting air quality index," *2021 from Innovation to Impact (FITI)*, 2021, . DOI: 10.1109/FITI54902.2021.9833033.

[38] T. M. Amado and J. C. Dela Cruz, "Development of machine learning-based predictive models for air quality monitoring and characterization," *TENCON 2018 - 2018 IEEE Region 10 Conference*, 2018, . DOI: 10.1109/TENCON.2018.8650518.

[39] O. Bouakline et al, "Prediction of daily PM10 concentration using machine learning," *2020 IEEE 2nd International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS)*, 2020, . DOI: 10.1109/ICECOCS50124.2020.9314380.

[40] A. Chauhan and P. R. Vamsi, "Anomalous ozone measurements detection using unsupervised machine learning methods," *2019 International Conference on Signal Processing and Communication (ICSC)*, 2019, . DOI: 10.1109/ICSC45622.2019.8938256.

[41] S. B. Kasetty and S. Nagini, "A survey paper on an IoT-based machine learning model to predict air pollution levels," *2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, 2022, . DOI: 10.1109/ICAC3N56670.2022.10074555.

[42] A. Catovic et al, "Air pollution prediction and warning system using IoT and machine learning," *2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME),* 2022, . DOI: 10.1109/ICECCME55909.2022.9987957.

[43] G. Gomathi et al, “Real time air pollution prediction in urban cities using deep learning algorithms and IoT,” *2022 7th International Conference on Communication and Electronics Systems (ICCES),* 2022, . DOI: 10.1109/ICCES54183.2022.9835991.

[44] M. Fahim et al, “A machine learning based analysis between climate change and human health: A correlational study,” *2022 International Conference on Computer and Applications (ICCA)*, 2022, . DOI: 10.1109/ICCA56443.2022.10039484.

[45] Y. -C. Lin et al, “Using machine learning to analyze and predict the relations between cardiovascular disease incidence, extreme temperature and air pollution,” *2021 IEEE 3rd Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (ECBIOS)*, 2021, . DOI: 10.1109/ECBIOS51820.2021.9510479.

[46] J. Collado and C. Pinzon, “Air pollution prediction using machine learning algorithms: A literature review,” *2022 V Congreso Internacional en Inteligencia Ambiental, Ingeniería de Software y Salud Electrónica y Móvil (AmITIC), San Jose, Costa Rica*, 2022, pp. 1-6, doi: 10.1109/AmITIC55733.2022.9941271.

[47] A. Dhumvad et al, “Water pollution monitoring and decision support system,” *2022 3rd International Conference for Emerging Technology (INCET)*, Belgaum, India, 2022, pp. 1-6, doi: 10.1109/INCET54531.2022.9824110.

[48] N. Rakesh and U. Kumaran, “Performance analysis of water quality monitoring system in IoT using machine learning techniques,” *2021 International Conference on Forensics, Analytics, Big Data, Security (FABS)*, Bengaluru, India, 2021, pp. 1-6, doi: 10.1109/FABS52071.2021.9702592.

[49] S. J. Sugumar et al, “Real time water treatment plant monitoring system using IOT and machine learning approach,” *2021 International Conference on Design Innovations for 3Cs Compute Communicate Control (ICDI3C)*, Bangalore, India, 2021, pp. 286-289, doi: 10.1109/ICDI3C53598.2021.00064.

[50] D. Jalal and T. Ezzedine, “Decision tree and support vector machine for anomaly detection in water distribution networks,” *2020 International Wireless Communications and Mobile Computing (IWCMC)*, Limassol, Cyprus, 2020, pp. 1320-1323, doi: 10.1109/IWCMC48107.2020.9148431.

[51] U. Shafi et al, “Surface water pollution detection using internet of things,” *2018 15th International Conference on Smart Cities: Improving Quality of Life Using ICT & IoT (HONET-ICT)*, Islamabad, Pakistan, 2018, pp. 92-96, doi: 10.1109/HONET.2018.8551341.

[52] S. Cao, S. Wang and Y. Zhang, “Design of river water quality assessment and prediction algorithm,” *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Orlando, FL, USA, 2018, pp. 901-906, doi: 10.1109/ICMLA.2018.00146.

[53] C. Lal and S. Kumar, “Ganga river water assessment using deep neural network: A study,” *2022 International Conference on Fourth Industrial Revolution Based Technology and Practices (ICFIRTP)*, Uttarakhand, India, 2022, pp. 184-186, doi: 10.1109/ICFIRTP56122.2022.10063185.

[54] K. Smolak et al, “Urban hourly water demand prediction using human mobility data,” *2018 IEEE/ACM 5th International Conference on Big Data Computing Applications and Technologies (BDCAT)*, Zurich, Switzerland, 2018, pp. 213-214, doi: 10.1109/BDCAT.2018.00036.

[55] A. N. Hasan and K. M. Alhammadi, “Quality monitoring of abu dhabi drinking water using machine learning classifiers,” *2021 14th International Conference on Developments in eSystems Engineering (DeSE)*, Sharjah, United Arab Emirates, 2021, pp. 1-6, doi: 10.1109/DeSE54285.2021.9719373.

[56] E. Kuruvilla and S. Kundapura, “Performance comparison of machine learning algorithms in groundwater potability prediction,” *2022 IEEE 7th International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, MANGALORE, India, 2022, pp. 53-58, doi: 10.1109/ICRAIE56454.2022.10054298.

[57] V. P. Parmar and A. J. Dhruv, “Efficient sea water purification using hybrid nanofiltration system and ML for optimization,” *2021 International Conference on Artificial Intelligence and Machine Vision (AIMV)*, Gandhinagar, India, 2021, pp. 1-6, doi: 10.1109/AIMV53313.2021.9670922.

[58] R. Gai and J. Yang, “Summary of water quality prediction models based on machine learning,” *2021 IEEE 23rd Int Conf on High Performance Computing & Communications; 7th Int Conf on Data Science & Systems; 19th Int Conf on Smart City; 7th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys)*, Haikou, Hainan, China, 2021, pp. 2338-2343, doi: 10.1109/HPCC-DSS-SmartCity-DependSys53884.2021.00353.

[59] Liming Zhang and Haowen Yan, “Implementation of a GIS-based water quality standards syntaxis and basin water quality prediction system,” *2012 International Symposium on Geomatics for Integrated Water Resource Management,* Lanzhou, 2012, pp. 1-4, doi: 10.1109/GIWRM.2012.6349656.

[60] F. R. Islam and K. A. Mamun, “GIS based water quality monitoring system in pacific coastal area: A case study for Fiji,” *2015 2nd Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE),* Nadi, Fiji, 2015, pp. 1-7, doi: 10.1109/APWCCSE.2015.7476226.

[61] M. G. Tyagunov and Z. Y. Lin, “Determining the optimal placements of renewable power generation systems using regional geographic information system,” *2017 2nd International Conference on the Applications of Information Technology in Developing Renewable Energy Processes & Systems (IT-DREPS),* Amman, Jordan, 2017, pp. 1-6, doi: 10.1109/IT-DREPS.2017.8277823.

[62] B. Pulido et al., “GIS-based DSS for optimal placement for oceanic power generation: OCEANLIDER project, Spanish coastline study,” 2013 International Conference on Renewable Energy Research and Applications (ICRERA), Madrid, Spain, 2013, pp. 137-142, doi: 10.1109/ICRERA.2013.6749740.

Estimación de incertidumbre para un sistema de reconocimiento de voz

Uncertainty estimation for a speech recognition system

Walter Morales-Muñoz[1], Saúl Calderón-Ramírez[2]

Morales-Muñoz, W; Calderón-Ramírez, S. Estimación de incertidumbre para un sistema de reconocimiento de voz. *Tecnología en Marcha*. Vol. 37, special issue. August, 2024. IEEE International Conference on BioInspired Processing. Pag. 97-103.

https://doi.org/10.18845/tm.v37i7.7305

[1] Instituto tecnológico de Costa Rica. Costa Rica.
wmorales@itcr.ac.cr
https://orcid.org/0000-0002-3888-4951

[2] Instituto Tecnológico de Costa Rica. Costa Rica.
sacalderon@itcr.ac.cr
https://orcid.org/0000-0001-9993-4388

Palabras clave

Incertidumbre; Reconocimiento de voz; ASR; Whisper; Monte Carlo Dropout.

Resumen

Whisper es un sistema de reconocimiento de voz diseñado por la compañía OpenAI, dicho sistema ha sido entrenado con 680,000 horas de datos supervisados multilingües y multitarea recopilados de la web. La siguiente investigación tiene como objetivo adaptar y emplear la técnica de Monte Carlo Dropout utilizando datos audios etiquetados en español y contaminados con una cantidad de ruido y la distancia de Levensthein para estimar la incertidumbre de dicho sistema. Resultados preliminares muestran que existe una relación lineal entre la estimación de la incertiumbre utilizando la distancia Levensthein y el medoide respecto al Word Error Rate (WER) de las transcripciones, además se observa que la cantidad de inserciones u omisiones en las transcripciones tiende a ser bajo.

Keywords

Uncertainty; Speech Recognition; ASR; Whisper; Monte Carlo Dropout.

Abstract

Whisper is a voice recognition system designed by the company OpenAI, which has been trained with 680,000 hours of multilingual and multitask supervised data collected from the web. The following research aims to adapt and employ the Monte Carlo Dropout using audio data labeled in Spanish and contaminated with a certain amount of noise and Levensthein distance to estimate the score uncertainty of this system.Preliminary results show that there is a linear relationship between uncertainty estimation and the Word Error Rate (WER) of the transcriptions. Furthermore, it is observed that the number of insertions or omissions in the transcriptions tends to be low.

Introducción

Los modelos de Deep Learning se centran en generar aplicaciones del mundo real, abarcando diversos ámbitos como la visión por computadora, el procesamiento del lenguaje natural, las finanzas, la robótica, el reconocimiento de voz y más. Independientemente del dominio de la aplicación, los modelos de Deep Learning utilizan redes neuronales profundas para aprender autónomamente cómo realizar tareas complejas a partir de grandes cantidades de datos. Para garantizar la fiabilidad de estos modelos en las distintas tareas específicas que realizan en aplicaciones cotidianas, es necesario evaluar cuán seguras son las predicciones que realizan estos modelos. Si los datos de prueba difieren significativamente de los datos de entrenamiento, es posible que el rendimiento del modelo no sea óptimo.

Como se menciona en [1], el tema de medir la calidad de los datos a través de diferentes métricas requiere más investigación, especialmente en el caso de datos no estructurados, que se utilizan comúnmente en la mayoría de las aplicaciones de Deep Learning. Este es el caso de aplicaciones que se centran en el reconocimiento automático de voz (ASR), donde las señales de audio (datos no estructurados) se utilizan como entrada para producir transcripciones precisas de voz a texto. Es importante comprender cómo se puede estimar la pérdida de confianza que el modelo puede experimentar en sus predicciones, ya sea debido a la variabilidad inherente en la adquisición de datos de entrada o a la arquitectura del modelo.

En el caso de los sistemas de Reconocimiento Automático de Voz (ASR), existen varios tipos, uno de los cuales es el sistema Whisper diseñado por la empresa OpenAI. Como se menciona en [2], este sistema ha demostrado ser uno de los sistemas ASR más potentes en la actualidad. Ha sido entrenado con 680,000 horas de datos supervisados multilingües recopilados de la web. Los autores afirman que el uso de un conjunto de datos tan grande y diverso conduce a una mayor robustez del sistema, incluso si los datos de entrada utilizados para las predicciones incluyen acentos específicos, ruido de fondo y otras variaciones.

Con esto en mente, el objetivo es evaluar la confiabilidad del sistema de reconocimiento de voz Whisper, donde se utilizará la técnica de Monte Carlo Dropout para capturar la incertidumbre del modelo. La importancia de resolver este problema radica en la necesidad de establecer técnicas que promuevan el uso seguro de aplicaciones impulsadas por IA, que ahora son ampliamente utilizadas por el público a diario. Esta investigación tiene como objetivo fomentar el uso seguro de varios modelos de acceso público y proporcionar ideas valiosas para empresas que dependen del sistema de reconocimiento Whisper y requieren un análisis detallado de su rendimiento.

La importancia de abordar este problema propuesto es evidente en un contexto en el que la Inteligencia Artificial ha experimentado un crecimiento exponencial, lo que ha llevado a la automatización de tareas que antes se consideraban imposibles de automatizar. Evaluar el rendimiento de estos modelos promueve la adopción segura de estas tecnologías.

En términos generales, para cualquier modelo de Deep Learning como Whisper, la incertidumbre del modelo, como se propone en [3], puede expresarse de acuerdo con la ecuación (1), y puede aproximarse como se muestra en la ecuación (2).

$$p(y^* \mid x^*, X, Y) = \int p(y^* \mid W, x^*, X, Y)\, p(W \mid x^*, X, Y)\, dW \tag{1}$$

$$\approx \int p(y^* \mid W, x^*)\, p(W \mid X, Y)\, dW \tag{2}$$

Según lo anterior, para calcular la incertidumbre de diferentes modelos de Deep Learning, es necesario modelar las distribuciones posteriores dadas por la ecuación (2). En términos generales, para cualquier modelo de Deep Learning como Whisper, la incertidumbre del modelo, como propone [3], puede expresarse de acuerdo con la ecuación (1), donde y^* es una posible transcripción dada por los parámetros W del modelo obtenidos a partir de un clip de audio de prueba x^* y los datos de entrenamiento previos X, Y. La ecuación (1) puede aproximarse como se muestra en la ecuación (2). Sin embargo, como afirma [4], para los modelos de Deep Learning, la distribución $p(W \mid X, Y)$ es intratable, y nuestro caso no es una excepción. La ecuación (4) depende de $p(Y \mid X)$, lo que significa que se basa en el cálculo de la verdadera distribución de probabilidad del conjunto de datos, y dada la naturaleza del problema, esto no se puede hacer analíticamente. Por lo tanto, se debe utilizar un enfoque para aproximar la ecuación (2). Hay varios enfoques informados en la literatura, pero un método comúnmente utilizado que no implica modificar la arquitectura del modelo es aproximar la distribución $p(W \mid X, Y)$ y así cuantificar la incertidumbre como se indica en la ecuación (2). Como sugiere [4], una manera de hacer esto es utilizando la aproximación mostrada en la ecuación (3) donde ϕ representa el dropout, un concepto ampliamente utilizado en aplicaciones de aprendizaje profundo. El término dropout puede entenderse como las tasas de Bernoulli que permiten la desactivación de neuronas específicas dentro de la arquitectura del modelo. El significado de dropout está determinado por la variable aleatoria de Bernoulli que decide si la entrada conectada en cada capa de la red neuronal debe ser descartada o no. Dado lo anterior, como se muestra en [4] y demostrado por [5], con la aproximación realizada, se puede mostrar que la incertidumbre epistémica del modelo, como se indica en la ecuación (2), puede aproximarse mediante (4). Donde $\{\mathrm{y}\}_{\mathrm{t}=1}^{\mathrm{T}}$ es un conjunto de *T* muestras de salida obtenidas de los parámetros *W* del modelo cuando se activa el dropout, y $\bar{y} = \frac{1}{T}\sum_{t=1}^{t} y_t$ es la media de la salida.

$$p(\, W \mid X, Y \,) \approx q(W; \Phi) = Bernoulli(W; \Phi) \tag{3}$$

$$p(\, y^* \mid x^*, X, Y \,) \approx Var_{p(\,y|x\,)}^{model}(y) = \sigma_{model} = \frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{y})^2 \tag{4}$$

Esta técnica es conocida como Monte Carlo Dropout, se utiliza dropout durante la inferencia para obtener una estimación de la incertidumbre epistémica del modelo y es una técnica muy utilizada en el área de visión computacional. Esta técnica se puede adecuar para problemas de reconocimiento de voz como veremos más adelante, la idea general es que el dropout se aplica múltiples veces en diferentes capas del modelo de red neuronal durante la inferencia. Al generar múltiples transcripciones para entradas de audio específicas, se obtienen múltiples transcripciones y múltiples probabilidades de selección de palabras, lo que conduce a una transcripción dada. Utilizando estas transcripciones, se puede calcular una medida de variabilidad de las transcripciones, proporcionando una así una medida de incertidumbre del modelo.

Es importante destacar que dado que el modelo predecirá una lista de palabras correspondientes al audio en cada iteración y, dada la arquitectura del decodificador basado en Transformer del modelo, como afirma [6], en cada iteración se pueden obtener transcripciones de longitudes variables. La forma en que se genera la estimación dependerá del procesamiento interno de las redes neuronales. Para generalizar la ecuación (4) para el caso del reconocimiento de voz y obtener una distribución de incertidumbre, se necesita una métrica de distancia no euclidiana d que pueda calcularse entre secuencias de caracteres con longitudes potencialmente diferentes. Como propone [6], la variabilidad en las transcripciones se puede obtener utilizando la distancia de Levenshtein, que es una métrica de distancia simple derivada del número de operaciones de edición requeridas para transformar una cadena de caracteres en otra. Dado que no hay noción de promedio en un espacio no euclidiano, se utilizará el medoide de las *T* diferentes transcripciones de salida y_{T} como nuestra “media” $\bar{y}$ de transcripciones, que puede escribirse como:

$$\bar{y} = \underset{y \in \{y_1, \dots y_T\}}{\operatorname{argmin}} \sum_t d(y, y_T)$$

Metodología

Se tomó una muestra de 10 archivos de audio del conjunto de datos CIEMPIESS-UNAM. La muestra fue contaminada con un 5% de ruido de fondo, mientras que el modelo Whisper se configuró con los siguientes parámetros: config.attention_dropout = 0.3, config.activation_dropout = 0.5, config.dropout = 0.5. Estos parámetros activan el dropout en diferentes capas de la arquitectura. Para cada uno de los archivos de audio, se utilizó el sistema Whisper (versión pequeña) para generar transcripciones, considerando 10 iteraciones por audio. Para cada una de las 10 iteraciones correspondientes a un audio específico, se calculó el medoide de las transcripciones y, posteriormente, se computó la distancia de Levenshtein al medoide encontrado. La función indicadora se utiliza para generar la probabilidad de obtener una distancia de Levenshtein específica. Esta probabilidad representa una forma de estimar numéricamente la incertidumbre en cuestión.

Para cada una de las transcripciones, se obtiene la tasa de error de palabras (Word Error Rate, WER) al comparar las transcripciones generadas con las transcripciones reales del audio. Esto permite vincular las incertidumbres obtenidas con una métrica comúnmente utilizada en sistemas de reconocimiento de voz.

Resultados

La Figura 1 muestra la dispersión de los datos. En el eje y se representa la distancia de Levenshtein con respecto al medoide (incertidumbre del puntaje) en función del WER; obtenido al comparar cada una de las transcripciones de audio con la verdad fundamental. Se observa un comportamiento lineal, como se esperaba, con un coeficiente de variación de R^2=0.62, lo que indica una correlación relativamente fuerte y positiva entre las dos variables representadas.

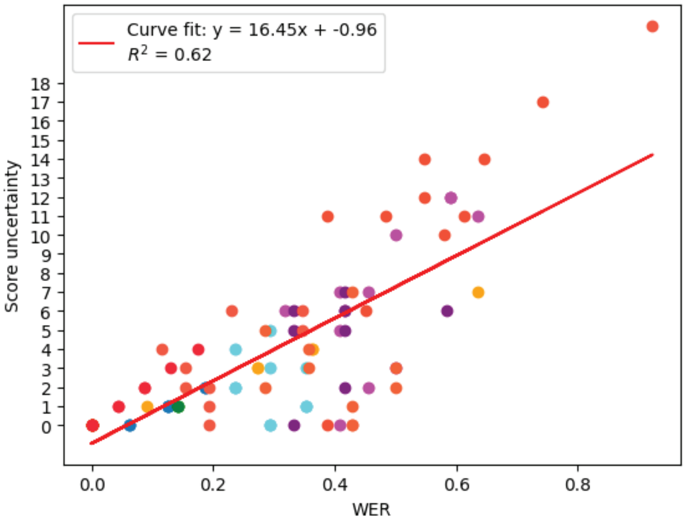


Figura 1. Puntaje de incertidumbre en función del WER.

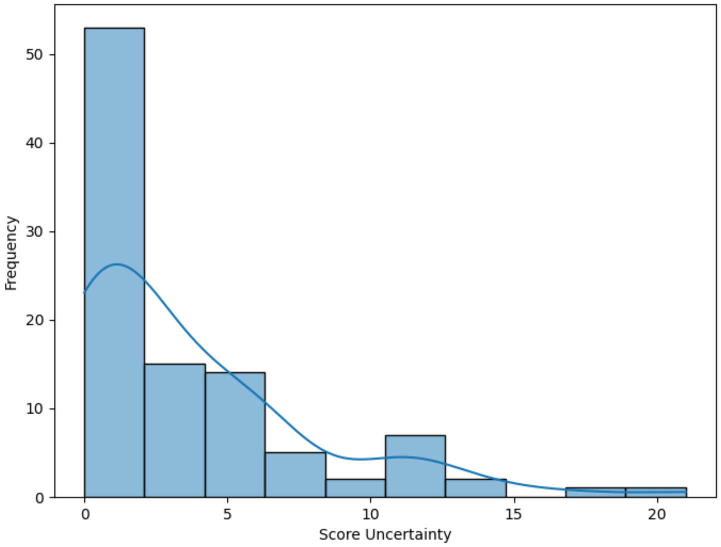


Figura 2. Distribución de frecuencia del puntaje de incertidumbre.

La Figura 2 muestra la distribución de frecuencia de la distancia de Levenshtein con respecto a los medoides de cada muestra de audio de prueba. Se puede observar que el pico del histograma está a la izquierda, lo que indica un bajo número de operaciones requeridas para convertir la cadena de texto de las transcripciones en el medoide correspondiente. La Figura 3 muestra las probabilidades de obtener una incertidumbre de puntuación específica dentro del conjunto obtenido. Se puede observar que la probabilidad más alta es 0.25, lo que corresponde a obtener un valor de distancia cero, lo que significa transcripciones sin variaciones en todas las iteraciones. Debido al comportamiento lineal observado previamente, si las transcripciones no varían en cada una de las iteraciones, se esperaría un WER cercano a cero, es decir, que la baja variabilidad en las predicciones resultaría en las transcripciones coincidiendo con la verdad fundamental.

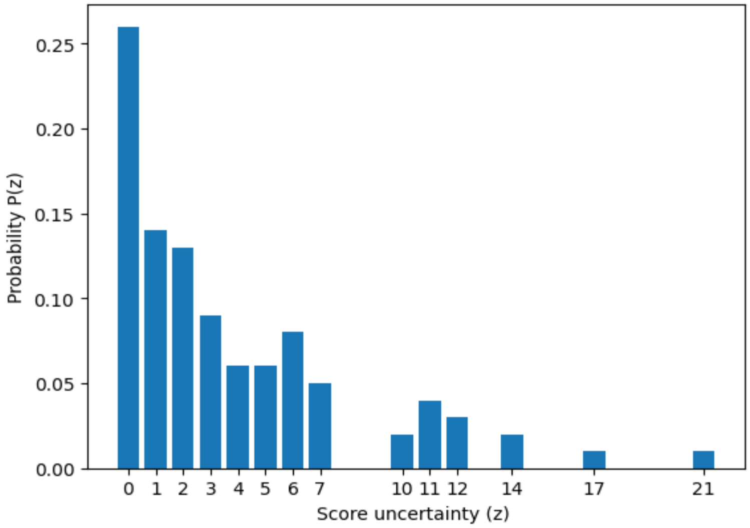


Figura 3. Distribución de frecuencia normalizada.

Conclusiones

Se propuso una técnica para estimar la incertidumbre de un modelo de reconocimiento automático de voz (ASR) basado en la arquitectura de un transformer. Para trabajos futuros, se recomienda aumentar el tamaño de la muestra, introducir variaciones en la configuración de dropout del modelo y utilizar muestras contaminadas con diferentes niveles de ruido, comparándolas con muestras no contaminadas, con el fin de generalizar los resultados mediante pruebas estadísticas e identificar cuantitativamente cómo el ruido afecta el rendimiento del modelo. Además, se sugiere obtener las probabilidades proporcionadas por el modelo para cada token seleccionado aplicando una función softmax a los logits de cada predicción de salida; con el fin de calcular una fiabilidad promedio del modelo y compararla con el WER generado en cada transcripción; si el modelo se comporta correctamente, se esperaría observar una pendiente negativa, lo que indica que un menor WER corresponde a una mayor fiabilidad en la salida generada por el modelo.

Referencias

[1] Díaz, C., Calderon-Ramirez, S., y Aguilar, L. D. M. (2022). Data quality metrics for unlabelled datasets. En 2022 ieee 4th international conference on bioinspired.

[2] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., y Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. arXiv preprint arXiv:2212.04356 .

[3] Mena, J., Pujol, O., y Vitria, J. (2021). A survey on uncertainty estimation in deep learning classification systems from a bayesian perspective. ACM Computing Surveys.

[4] Loquercio, A., Segu, M., y Scaramuzza, D. (2020). A general framework for uncertainty estimation in deep learning. IEEE Robotics and Automation Letters, 5 (2), 3153–3160.

[5] Gal, Y., y Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. En international conference on machine learning (pp. 1050–1059)

[6] Jayashankar, T., Roux, J. L., y Moulin, P. (2020). Detecting audio attacks on asr systems with dropout uncertainty. arXiv preprint arXiv:2006.01906.