Universidad de los Andes
Facultad de Ciencias
Departamento de Física
Laboratorio de Física Aplicada

Búsqueda de agrupaciones en data proveniente de electrocardiogramas (ECG), mediante el Análisis de Componentes Principales (PCA) y el uso de Redes Neuronales.

Trabajo especial de grado.

Br. Abrahan David Quintero Teran

Tutor: Prof. Juan Villegas

Jurados:
Prof. Marcos Rodríguez
Prof. John Ferreira

Mérida – Venezuela
2024

Resumen

Este es el resumen de mi tesis.

Índice general

Introducción

Capítulo 1

El problema

El electrocardiograma (ECG) es una técnica no invasiva que permite registrar y medir las señales eléctricas generadas por el corazón. Consiste en la captación de la variación temporal del potencial bioeléctrico durante cada ciclo cardíaco, utilizando electrodos colocados en la superficie cutánea del paciente [1]. El análisis del ECG proporciona datos sobre el sistema cardiovascular, en particular el corazón, lo que permite detectar diversas enfermedades que pueden afectar su funcionamiento óptimo. Estas enfermedades incluyen arritmias cardíacas, obstrucción de arterias, insuficiencia cardíaca y ataques al corazón [2].

Según la organización mundial de la salud [3], las enfermedades cardiovasculares (ECV) son las principal causa de muerte en hombres y mujeres en el mundo, con alrededor de 17,9 millones de personas que mueren al año a causa de estas. Entre los numerosos factores que llevan a esta consecuencia, se encuentran los errores provenientes de la interpretación manual de los electrocardiogramas (ECGs), por lo general, los médicos emplean características heurísticas diseñadas manualmente o utilizan arquitecturas de aprendizaje de características superficiales, esto puede generar como consecuencia, variabilidad entre los diagnósticos de los observadores e identificación de anomalías incorrectas que pueden llevar a ocasionar diagnósticos imprecisos y en consecuencia, tratamientos inadecuados. Además estos métodos manuales que utilizan arquitecturas de aprendizaje de características superficiales descartan información relevante del ECG inmersa dentro de características que no son superficiales, lo que provee una baja exactitud en el diagnostico a partir de las señales por lo que siempre será necesario de la supervisión de un experto con el fin de corregir estos errores.

1.1. Justificación

Ante la necesidad de reducir los errores provenientes de la interpretación manual de los ECGs que junto con la presencia de ruidos e interferencias en las señales complican aún más el análisis del ECG surge la necesidad de desarrollar herramientas que ayuden a reducir estos errores. Por lo que la identificación de patrones específicos y la clasificación precisa en grupos de pacientes siguen siendo desafíos importantes, ya que los métodos tradicionales de análisis de ECGs a menudo no son lo suficientemente robustos para manejar estas complejidades de manera eficiente. Es por esto que el desarrollo de herramientas que favorezcan la debida identificación de diversas enfermedades cardiovasculares, puede tener un impacto positivo en el proceso de diagnostico de enfermedades y facilitar tratamientos oportunos.

1.2. Antecedentes

Desde finales del siglo pasado ha habido un creciente interés en los métodos computacionales aplicados a la salud, por ejemplo en 1985, Pan y Tompkins [4] diseñaron un algoritmo para detección de los complejos QRS en tiempo real y con bajo costo computacional, este algoritmo sentó las bases de la implementación de algoritmos de detección y análisis para las señales de electrocardiogramas. En 1996, Laguna *et al.* [5] presentaron un sistema de estimación del modelo de Hermite adaptativo (AHMES) para la estimación en línea latido a latido de las características que describen el complejo QRS con el modelo de Hermite. Gómez Herrero *et al.* [6], presentaron un algoritmo conocido como "Matching Pursuit" que ofrece la capacidad de descomponer cualquier señal en un combinación lineal de formas de onda extraídas de un diccionario redundante de funciones llamado Gabor. Este algoritmo se ha reconocido como una herramienta eficaz para realizar transformaciones adaptativas de tiempo-frecuencia en señales de ECG, lo que permite obtener características relevantes en el dominio tiempo-frecuencia.

Siguiendo esta linea de algoritmos que realizan transformaciones adaptativas en el dominio tiempo-frecuencia, se encuentran Martínez *et al.* [7] quienes proponen un delineador de ECG basado en las transformadas Wavelet (TW), para así poder detectar los inicios, picos y finales de las ondas P y T como también las ondas individuales del complejo QRS incluyendo el inicio y fin del complejo, todo esto para luego determinar los diversos intervalos y segmentos dentro del ECG, este método propuesto es tan robusto que no se ve afectado por los

diversos ruidos que pueden existir dentro del ECG como lo es por ejemplo, el desplazamiento de la linea base. Para remover este ultimo, Sharma y Sharma en 2015 [8] usan la Descomposición Vibracional de Hilbert (HVD) para descomponer la señal original del ECG en una serie de funciones modales intrínsecas y luego remover el primer termino, que corresponde a la componente de mayor energía y así eliminar el desplazamiento de la linea base del ECG.

Otros métodos relevantes usando para la extracción de características a partir de dominios transformados, tales como la Transformada de Coseno Discreta (DCT), la Transformada Wavelet Continua (CWT) y la Transformada Wavelet Discreta (DWT), estas técnicas permiten analizar las señales de ECG en diferentes representaciones y obtener características significativas para su posterior procesamiento y análisis, así Khorrami y Moavenian en 2010 [9] utilizaron las CWT, DWT y DCT, con el fin de mejorar la capacidad de dos clasificadores de patrones en la clasificación de arritmias ECG.

Song *et al.* (2005) [10] extrajeron diecisiete características de entrada originales de señales pre-procesadas mediante TW, utilizando el análisis discriminaste lineal (LDA). El rendimiento del clasificador SVM (Máquina de Soporte Vectorial) con características reducidas por LDA mostró ser mayor que con el Análisis de Componentes Principales (PCA) e incluso con características originales, sin embargo, está técnica requirió de un mayor costo computacional. Yu y Chen (2007) [11] utilizaron la transformación wavelet y una red neuronal probabilística (PNN), para descomponer las señales de latido de ECG en diferentes sub-bandas utilizando la DWT. Posteriormente, seleccionaron tres conjuntos de características estadísticas de las señales compuestas para caracterizar las señales de ECG, así como la potencia de AC y el intervalo RR instantáneo de la señal original.

Ye, Coimbra y Kumar (2010) [12] propusieron un enfoque de combinación de características morfológicas y dinámicas refiriéndose a la TW y el análisis de componentes independientes (ICA) aplicándose por separado a cada latido del corazón para extraer los coeficientes correspondientes como características morfológicas. Además concatenaron la información del intervalo RR y estos dos tipos diferentes de características y se utilizó SVM para la clasificación. Rojas, Medina y Dugarte (2011) [13] diseñan un sistema multicanal de adquisición y analizan la señal electrocardiográfica de alta resolución, en el que utilizan Máquinas de Soporte Vectorial de mínimos cuadrados (LSSVM) para determinar el inicio del complejo QRS y el final de la onda T, entrenadas en base a atributos extraídos de la señal preprocesada y de señales obtenidas mediante

descomposiciones con Wavelets. Estas técnicas permiten estimar el intervalo QT así como el intervalo QT corregido (QTc).

También Zhang *et al.* (2024) [14] proponen un modelo de Redes Neuronales Convolucionales (CNN) para clasificar insuficiencias cardíacas por clases, según la Asociación del Corazón de Nueva York, a partir de imágenes electrocardiográficas en el que consiguen el mejor resultado segmentando las imágenes en fragmentos de 12 segundos. Así también Astudillo *et al.* (2024) [15] prueban cinco arquitecturas de CNN diferentes para clasificación de arritmias cardíacas. Estas dos ultimas investigaciones consiguen predicciones superiores al 98.98 % en el mejor de los escenarios.

En este sentido, Pan y Tompkins obtuvieron un 99.3 % de complejos QRS detectados haciendo uso de la base de datos MIT-BIH Arritmia (MITDB) [4]. El sistema que Laguna *et al.* [5] presentan, mejora la relación señal-ruido (SNR) en la estimación, lo que permite la adaptación a los cambios del QRS latido a latido, proporcionando una descripción de la evolución de la señal QRS y la compresión de datos del ECG. El sistema AHMES permite la estimación en línea de estas características con una mejor SNR que la estimación directa. Martínez *et al.* [7] usando MITDB encuentran un 99.8 % de complejos QRS detectados, un resultado superior al obtenido por Pan y Tompkins. Gómez Herrero *et al.* [6] usando la base de datos MITDB e introduciendo Análisis de Componentes Independientes (ICA) como extractor de características para el procesamiento del ECG para la simulación, obtienen resultados con una precisión de 99.8 % para clasificar latidos que corresponden a la clase de Ritmo Sinusal Normal (RSN) y 97.9 % para latidos de la clase de Contracción Ventricular Prematura (PVC). Así también Sharma y Sharma [8] compara su método usando criterios de correlación y SNR donde concluyen que la técnica propuesta se desempeña mejor en la mayoría de los casos que técnicas anteriores para remover el desplazamiento de la linea base. Song *et al.* [10] identificaron seis tipos diferentes de arritmias obteniendo una exactitud del 98.94 %. Ye, Coimbra y Kumar [12] reconocieron 15 clases de latidos con una exactitud del 99.66 % en un grupo de prueba de 85945 muestras. Khorrami y Moavenian [9], utilizando SVM y la base de datos MITDB con dos conjuntos de datos de prueba con diferentes configuraciones obtuvieron un error cuadrático medio (MSE) de 0.14 y 0.15 respectivamente para cada prueba.

Rojas, Medina y Dugarte [13] encuentran diferencias estadísticas significativas importantes entre pacientes chagásicos y pacientes de control de esta manera logran abordar la detección temprana y no invasiva de enfermedades cardio-

vasculares como el mal de Chagas.

1.3. Objetivos

Objetivo General

Analizar los datos provenientes de electrocardiogramas mediante la aplicación de análisis de componentes principales (PCA) y redes neuronales, con el objetivo de identificar agrupaciones en dicha data.

Objetivos específicos

1. Analizar y catalogar los datos de los electrocardiogramas para mejorar la calidad de los datos y la precisión de los modelos predictivos.

2. Desarrollar un marco metodológico solido que combine el análisis de componentes principales (PCA) y redes neuronales para la identificación de grupos de datos.

3. Evaluar el rendimiento de los modelos propuestos utilizando los datasets de Physionet.

Capítulo 2

Marco Teórico

2.1. Conceptos básicos

Introducción al Electrocardiograma (ECG)

La humanidad siempre, en su continuo deseo de aprender y entender más, ha querido desentrañar los secretos del cuerpo humano entendiendo su funcionamiento interno. Al principio con técnicas invasivas acordes a la tecnología disponible a la época, pero evolucionando continuamente, creando así exámenes cada vez menos invasivos, con el fin de mejorar el diagnostico, siendo mas preciso y oportuno. Entre esos exámenes se destaca el electrocardiograma (ECG) el cual es una representación visual de la actividad eléctrica del corazón en función del tiempo, que se obtiene desde la superficie corporal, con un electrocardiógrafo, este es el instrumento principal de la electrofisiología cardíaca y tiene una función relevante en el cribado y diagnóstico de las enfermedades cardiovasculares, alteraciones metabólicas y demás utilidades.

El primer «electrograma» humano fue publicado en 1887 por el fisiólogo británico Augustus Desiré Waller, de la St. Mary's Medical School de Londres. Utilizó un electrómetro capilar de Lipmann con electrodos aplicados a la espalda y el tórax del sujeto. Demostró que la contracción ventricular precedía a la actividad eléctrica. En su primer informe sobre un registro de la electricidad cardíaca realizado en la superficie corporal, Waller utilizó el término «cardiógrafo».

Einthoven empezó a experimentar con el potencial del capilar para captar corrientes eléctricas diminutas. En 1895 demostró cinco deflexiones que denominó

ABCDE en 1895. Creó un ajuste matemático para para tener en cuenta la inercia del sistema capilar, lo que produjo las curvas de corriente que vemos hoy en día. Siguiendo la tradición matemática establecida por Descartes, utilizó la parte terminal de la serie parte terminal de la serie alfabética (PQRST) para denominar estas derivaciones.

El pionero de la electrocardiografía, Waller dijo a finales de 1911: «No creo que la electrocardiografía vaya a tener un uso extensivo en los hospitales. A lo sumo puede tener un uso raro y ocasional para proporcionar un registro de alguna anomalía de la actividad cardíaca». Sin embargo, diez años de los estudios clínicos de Einthoven con los galvanómetros de cuerda transformaron este curioso fenómeno fisiológico en un dispositivo de registro clínico indispensable. Las asociaciones de la inversión de la onda T con la angina de pecho y la arteriosclerosis. en 1910, junto con otras arritmias, como el bigeminismo, bloqueo cardíaco completo, hipertrofia ventricular derecha e izquierda, fibrilación y aleteo auricular y ejemplos de diversas cardiopatías. Con su nueva técnica, estandarizó los trazados y formuló el concepto de «triángulo de Einthoven» relacionando relacionando matemáticamente las 3 derivaciones (Derivación III = Derivación II - Derivación I). En 1924, el «Padre de la electrocardiografía» recibió el Premio Nobel de Medicina [16].

En 1957, el médico estadounidense Norman Jefferis Holter inventó el ECG dinámico (DCG), a menudo conocido como Holter, en uno de los primeros intentos de combinar monitorización clínica y movilidad. Creó una mochila que pesaba unos 38 kg y tenía un dispositivo que podía registrar la actividad cardíaca del participante. Este portátil permite la monitorización continua de actividad eléctrica del sistema cardiovascular durante 24 horas, lo que ayuda a estudiar las arritmias y a localizar el lugar de la isquemia miocárdica. Reconociendo los beneficios potenciales de un dispositivo de monitorización de este tipo, Holter consiguió convertir su idea en una valiosa herramienta de diagnóstico reduciendo el tamaño y el peso a 1 kg con ayuda de Del Mar Avionics, un conocido fabricante de equipos aeronáuticos [16].

Durante las tres primeras décadas del siglo 20, el ECG de tres derivaciones periféricas fue largamente usado, especialmente luego de mejoras que lo hicieron más portable. A pesar de que el ECG de tres derivaciones era una manera fiable de evaluar arritmias, pronto se reconoció que el corazón incluía «zonas silenciosas» en las que un infarto de miocardio podría no ser detectado. En 1942, Emanuel Goldberger construyó las derivaciones precordiales (unipolares) usando el promedio de las diferencias de potencial de las tres derivaciones per-

iféricas como terminal de referencia, que inicialmente fue creado por Frank N. Wilson, al cual se le conoce como terminal central de Wilson, que ahora se denominan como derivaciones precordiales (V1-V6), donde en 1938, la Asociación Americana del Corazón (AHA) y la Sociedad Cardiaca de Gran Bretaña recomendaron la estandarización del posicionamiento de los electrodos en el pecho para dichas derivaciones. También Goldberger propuso una manera de obtener lo que ahora se llaman derivaciones aumentadas, conocidas por las siglas a-VL, a-VR, y a-VF. 8 años después la AHA recomendó la estandarización del ECG de 12 derivaciones. [16]

En la era digital, la tecnología del silicio y los circuitos circuitos impresos han hecho posible la miniaturización electrónicos. Desde hace algún tiempo, la tecnología ha ganado popularidad en el campo de la medicina y la necesidad de los clientes de controlar su salud ha sido el principal motor. La influencia de los "vestibles"(wearables) ha hecho inevitable la continua investigación y desarrollo de nuevas funciones que pueden evaluar y transmitir datos biométricos en tiempo real.

El corazón consta de cuatro partes, dos aurículas y dos ventrículos, el ECG registra los impulsos eléctricos que estimulan estas partes y producen su contracción. Las células cardíacas en reposo se encuentras cargadas o polarizadas; pero la estimulación eléctrica las "despolariza", y se contraen.

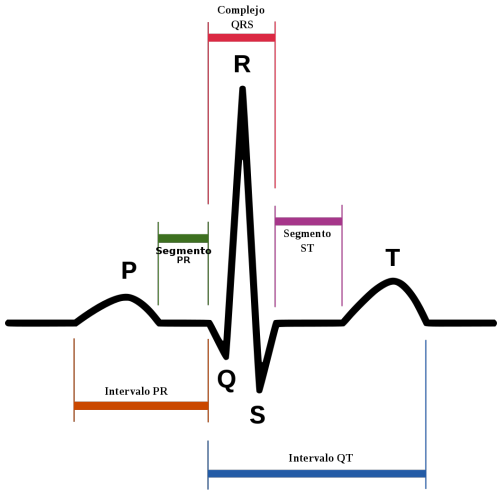

Figura 2.1: ECG del corazón con ritmo sinusal normal

Como se observa en la figura 2.1 el ECG consta de varias ondas representativas de cada etapa de un latido cardíaco, estas son:

- **Onda P**: registra la despolarización auricular.
- **Complejo QRS**: Es la despolarización ventricular.
- **Onda T**: representa la repolarización ventricular.

2.2. Conceptos específicos

Análisis de Componentes Principales (PCA).

Los grandes conjuntos de datos están cada vez más extendidos en muchas disciplinas. Para interpretarlos, se necesitan métodos que reduzcan drásticamente su dimensionalidad de forma interpretable, de modo que se conserve la mayor parte de la información contenida en los datos. Se han desarrollado muchas técnicas con este fin, pero el análisis de componentes principales (PCA) es una de las más antiguas y utilizadas. Su idea es sencilla: reducir la dimensionalidad de un conjunto de datos conservando la mayor cantidad posible de «variabilidad» (es decir, de información estadística).

Esto significa que «preservar tanta variabilidad como sea posible» se traduce en encontrar nuevas variables que sean funciones lineales de las del conjunto de datos original, que maximicen sucesivamente la varianza y que no estén correlacionadas entre sí. Encontrar esas nuevas variables, los componentes principales (PC), se reduce a resolver un problema de auto-valores y auto-vectores.

Hasta que no se generalizó el uso de ordenadores electrónicos, que fue posible utilizarlo con conjuntos de datos que no fueran trivialmente pequeños. Desde entonces, su uso se ha multiplicado y se han desarrollado numerosas variantes en muchas disciplinas diferentes. Desde entonces, su uso se ha multiplicado y se han desarrollado numerosas variantes en muchas disciplinas diferentes.

La definición formal de PCA, en un contexto estándar, junto con una derivación que muestra que puede obtenerse como la solución a un problema de auto-valores y auto-vectores o, alternativamente, a partir de la descomposición del valor singular (SVD) de la matriz (centrada) de datos. El PCA puede basarse en la matriz de covarianzas o en la matriz de correlaciones. Se discutirá la elección entre estos análisis. En ambos casos, las nuevas variables (las PC) dependen del conjunto de datos, en lugar de ser funciones de base predefinidas,

por lo que son adaptativas en sentido amplio. Los principales usos del PCA son descriptivos y no inferenciales.[17]

Redes Neuronales

Una red neuronal artificial es un grupo de neuronas artificiales interconectadas que interactúan entre sí de forma concertada. Se trata de hecho un procesador distribuido masivamente paralelo que tiene una propensión natural a almacenar el conocimiento experiencial y ponerlo disponible para su uso. Se parece al cerebro humano en dos aspectos: La red adquiere los conocimientos mediante un proceso de aprendizaje. El conocimiento lo adquiere la red mediante un proceso de aprendizaje, y para almacenarlo se utilizan las intensidades de conexión interneuronal, denominadas pesos. Los modelos de redes neuronales artificiales pueden utilizarse como método alternativo en análisis y predicciones. Funcionan como un modelo de «caja negra», que no requiere información detallada sobre el sistema. Imitan en cierto modo el proceso de aprendizaje de un cerebro humano porque aprenden la relación entre los parámetros de entrada, las variables controladas y no controladas estudiando datos registrados previamente. En este sentido, funcionan de forma similar a la regresión no lineal, pero son mucho más potentes que el análisis de regresión. Las redes neuronales son capaces de manejar sistemas grandes y complejos con muchos parámetros interrelacionados. Parece que simplemente ignoran el exceso de parámetros de entrada que tienen una importancia mínima y se concentran en los más importantes.

K-means

Capítulo 3

Marco experimental

...

Capítulo 4

Resultados

...

Parte I

Conclusiones

Capítulo 5

Conclusiones

...

Bibliografía

[1] L. Zhang, M. Karimzadeh, M. Welch, C. McIntosh, and B. Wang, "Chapter 7 - analytics methods and tools for integration of biomedical data in medicine," in *Artificial Intelligence in Medicine* (L. Xing, M. L. Giger, and J. K. Min, eds.), pp. 113–129, Academic Press, 2021.

[2] M. National Library of Medicine, "Electrocardiogram," 2020.

[3] O. Organizacion Mundial de la Salud, "Enfermedades cardiovasculares (cvds)," 2021.

[4] J. Pan and W. J. Tompkins, "A real-time qrs detection algorithm," *IEEE Transactions on Biomedical Engineering*, vol. BME-32, no. 3, pp. 230–236, 1985.

[5] P. Laguna, R. Jané, S. Olmos, N. Thakor, H. Rix, and P. Caminal, "Adaptive estimation of qrs complex wave features of ecg signal by the hermite model," *Medical and biological, engineering and computing*, vol. 34, pp. 58–68, 02 1996.

[6] G. Herrero, A. Gotchev, I. Christov, and K. Egiazarian, "Feature extraction for heartbeat classification using independent component analysis and matching pursuits," in *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 4, pp. iv/725–iv/728 Vol. 4, 2005.

[7] J. P. Martínez, R. Almeida, S. Olmos, A. P. Rocha, and P. Laguna, "A wavelet-based ecg delineator: Evaluation on standard databases," *IEEE transactions on bio-medical engineering*, vol. 51, pp. 570–81, 05 2004.

[8] H. Sharma and K. Sharma, "Baseline wander removal of ecg signals using hilbert vibration decomposition," *Electronics Letters*, vol. 51, pp. 447–449, 03 2015.

[9] H. Khorrami and M. Moavenian, "A comparative study of dwt, cwt and dct transformations in ecg arrhythmias classification," *Expert Syst. Appl.*, vol. 37, pp. 5751–5757, 08 2010.

[10] M. Song, J. Lee, S. Cho, K.-J. Lee, and S. Yoo, "Support vector machine based arrhythmia classification using reduced features," *International Journal of Control, Automation and Systems*, vol. 3, 12 2005.

[11] S.-N. Yu and C. Hsiang, "Electrocardiogram beat classification based wavelet and probabilistic neural network," *Pattern Recognition Letters*, vol. 28, pp. 1142–1150, 07 2007.

[12] C. Ye, M. Coimbra, and B. Kumar, "Arrhythmia detection and classification using morphological and dynamic features of ecg signals," *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, vol. 2010, pp. 1918–21, 08 2010.

[13] N. Dugarte Jerez, R. Medina Molina, and R. Rojas Sulbarán, "Sistema para la adquisición de la señal electrocardiógrafica de alta resolución," *Universidad, Ciencia y TecnologÃa*, vol. 15, pp. 206 – 215, 12 2011.

[14] C.-J. Zhang, Yuan-Lu, F.-Q. Tang, H.-P. Cai, Y.-F. Qian, and Chao-Wang, "Heart failure classification using deep learning to extract spatiotemporal features from ecg," *BMC Medical Informatics and Decision Making*, vol. 24, 01 2024.

[15] V. Astudillo, D. Luna, and J. Muñoz Chaves, "Clasificación de arritmias cardiacas usando redes neuronales convolucionales en muestras de ecg," *Revista EIA*, vol. 21, 01 2024.

[16] R. Vincent, "From a laboratory to the wearables: a review on history and evolution of electrocardiogram," *Iberoamerican Journal of Medicine*, vol. 4, no. 4, pp. 248–255, 2022.

[17] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments.," *Philosophical transaction. Series A, Mathematical, physical, and engineering sciences*, 2016.