Modelo de detección de cáncer cervical en muestras de tejido celular utilizando Máquinas de Soporte Vectorial

Daniel Alberto López Sánchez

July Andrea Muñoz Lopera

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesora

María Bernarda Salazar Sánchez, Doctor (PhD) en Ingeniería Electrónica

Universidad de Antioquia

Facultad de Ingeniería

Especialización en Analítica y Ciencia de Datos

Medellín, Antioquia, Colombia

2023

Cita	(López Sánchez & Muñoz Lopera, 2023)
Referencia **Estilo APA 7 (2020)**	López Sánchez, D. A., & Muñoz Lopera, J. A. (2023). *Análisis de cáncer cervical en diferentes tipos de muestras de tejido celular utilizando técnicas de aprendizaje automático* [Trabajo de grado especialización]. Universidad de Antioquia, Medellín, Colombia.

Especialización en Analítica y Ciencia de Datos, Cohorte IV.

Grupo de Investigación Intelligent Information Systems Lab.

Centro de Investigación Ambientales y de Ingeniería (CIA).

Dedicatoria

A nuestras familias por la paciencia y acompañamiento.

Agradecimientos

A los profesores por impartirnos su conocimiento y guiarnos en el proceso.

Tabla de contenido

Lista de tablas

Lista de figuras

Resumen

En esta monografía se genera un modelo de detección de cáncer cervical usando máquinas de soporte vectorial para clasificar diferentes tipos de células basados en sus características citomorfológicas.

Para lo anterior se establece una metodología de varias etapas. Primero se comienza con el preprocesamiento y preparación de los datos, en donde se extraen características de las imágenes en cada canal de color RGB. Se utilizan siete características: intensidad, suavidad, uniformidad, tercer momento, entropía, desviación estándar y mediana. Además, se usa la característica luminancia en escala de grises. Con lo anterior se hace una detección y eliminación datos atípicos utilizando el método LOF (Local Outlier Factor).

Luego, se procede a evaluar diferentes técnicas de escalamiento: Robusto, Estándar y Min-Max, y de reducción de dimensionalidad: Análisis de Componentes Independientes (ICA), Análisis de Componentes Principales (PCA) y Análisis Discriminante Lineal (LDA). Con estas técnicas se busca simplificar la representación de los datos y evitar el sobreajuste.

En la aplicación del modelo de máquinas de soporte vectorial (SVM del inglés Support Vector Machine), se utiliza el mejor método de reducción de dimensionalidad obtenido previamente y se evalúa el modelo mediante la exactitud ajustando diferentes hiper parámetros como el kernel, la regularización C, el valor gamma, el coef0 y el degree, buscando obtener el mejor rendimiento del modelo.

Palabras clave: células, cáncer cervical, detección, enfermedades, modelo, extracción de características, extracción datos atípicos, reducción dimensionalidad, máquinas de soporte vectorial, hiper parámetros, LOF, ICA, SVM, t-SNE.

Abstract

In this monograph a cervical cancer detection model is generated using support vector machines to classify different cell types based on their cytomorphological characteristics.

A multi-stage methodology is established for this purpose. First, we start with the preprocessing and preparation of the data, where features are extracted from the images in each RGB color channel. Seven features are used: intensity, smoothness, uniformity, third moment, entropy, standard deviation and median. In addition, the grayscale luminance feature is used. This is used to detect and eliminate outliers using the LOF (Local Outlier Factor) method.

Then, different scaling techniques are evaluated: Robust, Standard and Min-Max, and dimensionality reduction: Independent Component Analysis (ICA), Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). These techniques seek to simplify the representation of the data and avoid overfitting.

In the application of the SVM model, the best dimensionality reduction method previously obtained is used and the model is evaluated through accuracy by adjusting different hyper parameters such as kernel, C regularization, gamma value, coef0 and degree, seeking to obtain the best performance of the model.

Keywords: cells, cervical cancer, detection, diseases, model, feature extraction, outlier extraction, dimensionality reduction, support vector machines, hyper parameters, LOF, ICA, SVM, t-SNE.

1. Introducción

El cáncer cervical es una enfermedad grave que afecta a miles de mujeres en todo el mundo. A pesar de los avances en la detección y tratamiento, el cáncer cervical sigue siendo una de las principales causas de mortalidad femenina. La detección temprana es clave para su tratamiento exitoso, pero esta detección, por mucho tiempo ha dependido en gran medida del examen de *Papanicolaou,* una herramienta estándar para la detección de células anormales en el cuello uterino (American Cancer Society, 2021). Este examen de lectura y análisis de las muestras es considerado subjetivo, ya que su precisión puede variar según varios factores, como la calidad de la muestra, la habilidad del profesional médico que realiza la prueba y el estado de la enfermedad. Los diferentes estudios individuales sobre pruebas de Papanicolaou convencionales presentan una amplia variabilidad en las estimaciones de sensibilidad y especificidad, por ejemplo, registran estudios con sensibilidades entre el 30.0% y el 87.0%, mientras que la especificidad varía entre el 86.0% y el 100% (Nanda et al., 2000).

En los últimos años, el aprendizaje automático ha emergido como una técnica prometedora para analizar grandes cantidades de datos de muestras de tejido celular y para la creación de modelos de predicción más precisos y accesibles gracias al aumento de la disponibilidad de datos y al poder de procesamiento de las computadoras modernas. Se ha demostrado en varios estudios, como se detalla en el estado del arte, que ciertos algoritmos de inteligencia artificial son capaces de detectar y contar con gran precisión la presencia de diversos tipos de células en muestras de tejido cervical. Estas herramientas tecnológicas representan una alternativa altamente efectiva en comparación con los métodos tradicionales de conteo manual, que apoyan el posterior diagnóstico que realiza el profesional experto, dando sentido a la información procesada y a la decisión final del diagnóstico brindado al paciente.

Es evidente entonces la necesidad de desarrollar herramientas de apoyo al diagnóstico que aumenten la precisión y disminuyan la subjetividad en la detección temprana del cáncer cervical a partir de muestras de tejido del cuello uterino. Es así como en este proyecto se propone un modelo de detección de cáncer cervical desarrollado utilizando Máquinas de Soporte Vectorial (SVM del inglés Support Vector Machine), a partir del análisis de imágenes de muestras de tejido celular disponibles en SIPaKMeD (SIPaKMeD, 2020).

2. Planteamiento del problema

El cáncer cervical es una de las enfermedades más comunes y letales en mujeres a lo largo del mundo, afectando a un total de 500,000 y causó solo 300,000 muertes en 2021 (Manna et al., 2021). Sumado a esto, su detección suele ser compleja ya que requiere que los profesionales clasifiquen cada célula de muestras de tejido cervical de más de 100,000 células con el método de Papanicolaou. Este procedimiento intenso y costoso limita la detección de cáncer cervical en la población, principalmente en países en vía de desarrollo, donde las tasas de detección temprana y tratamiento son bajas. En la actualidad, el diagnóstico de cáncer cervical se realiza mediante la evaluación de muestras de tejido celular obtenidas por biopsia, lo que requiere la intervención de un especialista (American Cancer Society, 2021).

En Colombia, el cáncer cervical también es un problema importante de salud pública. Según el Ministerio de Salud y Protección social, este es el segundo tipo de cáncer entre las mujeres colombianas y es la primera causa de muerte por cáncer entre mujeres de 30 a 59 años, cada día 12 mujeres son diagnosticadas con esta enfermedad y 5 colombianas mueren diariamente por esta patología. A esto se le suma que la mortalidad por cáncer de cuello uterino se asocia a condiciones socioeconómicas desfavorables, encontrándose un mayor riesgo de mortalidad en regiones rurales dispersas, con bajo acceso a los servicios de salud y en grupos de menor nivel educativo (Ministerio de Salud y Protección Social de Colombia, 2021). A pesar de los programas de detección y prevención existentes, la tasa de mortalidad por cáncer cervical sigue siendo alta, sobre todo para las regiones con menos acceso a profesionales de la salud.

Aún con todos los avances en la tecnología médica y la mejora de las técnicas de diagnóstico, la detección temprana de esta enfermedad sigue siendo un desafío. Aunque existen pruebas para la detección de células anormales en el cuello uterino, la sensibilidad y especificidad de las pruebas actuales son limitadas y los resultados a menudo son subjetivos, lo que puede llevar a errores de diagnóstico y retrasos en el tratamiento (American Cancer Society, 2021). Para mejorar esta situación, algunos profesionales se han ayudado de los desarrollos en tecnologías de aprendizaje automático (ML, del inglés Machine Learning), los cuales han presentado buenos resultados en temas relacionados con clasificación y detección del cáncer, como por ejemplo la

técnica de optimización del lobo gris (GWO) con una precisión del 98.1% en el proceso de clasificación (Basak et al., 2021) y modelos de aprendizaje de transferencia, con rendimientos del 96.9% (Pramanik et al., 2022).

Actualmente la mayor parte de los desarrollos en cáncer cervical se han enfocado en la clasificación de imágenes para detectar el tipo de células presentes. Sin embargo, dada la necesidad de contar el número de células anormales en las muestras de tejido, como apoyo al diagnóstico, puede ser útil un modelo que le permita al profesional apoyar su análisis de la muestra con una segunda lectura antes de hacer el diagnóstico final.

Para abordar este problema, en este trabajo se propone desarrollar un modelo de detección de cáncer cervical en muestras de tejido celular utilizando la técnica de aprendizaje automático de SVM. El modelo utilizará datos clínicos y características histológicas de las muestras para predecir la presencia o ausencia de cáncer cervical con alta precisión, tal que con una exactitud del 95.0% realice de forma temprana y rápida la interpretación de la muestra de tejido de celular.

3. Justificación

Como ya se ha mencionado previamente, la detección temprana del cáncer cervical es crucial para prevenir el desarrollo de la enfermedad y mejorar las tasas de supervivencia, por ende la creación de un modelo de predicción aplicable, utilizando técnicas de aprendizaje automático, toma relevancia debido a la alta incidencia de cáncer cervical a nivel mundial, más aun teniendo en cuenta que en países de bajos y medianos ingresos, la incidencia y la mortalidad son aún más altas debido a la falta de programas de detección y tratamiento adecuados.

Si a lo anterior se le suma que la tasa relativa de supervivencia a 5 años del cáncer cervical es del 67.0%, y qué esta tasa es dependiente del estado de diseminación del cáncer al momento de ser diagnosticado, tal que, en estado temprano es del 92.0%, cuando ha diseminado a tejidos u órganos cercanos alcanza un 59.0% y en la etapa tardía de diagnóstico, cuando el cáncer ha afectado partes lejanas del cuerpo, la tasa puede ser tan baja como un 17.0% (National Cancer Institute, 2021).

Sin contar con que la tasa de incidencia y mortalidad puede verse agravada por el nivel de desarrollo de los países, por ejemplo, la tasa de incidencia por 100,000 mujeres en Canadá es de 6.3 y en Estados Unidos es de 6.6 con tasas de mortalidad de 1.7 y 2.7 respectivamente, en contraste con incidencias de 32.7 en Perú, 32.8 en Venezuela, 36.2 en Nicaragua y 23.3 en México, y una mortalidad de 12.0, 12.3, 18.3 y 8.0 respectivamente, datos reportados por la Organización Panamericana de la Salud (OPS) y la Organización Mundial de la Salud (OMS) (Pérez-González & Aguilar-Lemarroy, 2015). Solo en Colombia, según el DANE, para el año 2022 se tuvo dentro de la población rural 5.9 millones de mujeres (DANE, 2022), siendo esta población la que presenta mayores dificultades para acceder a servicios de salud y especialistas médicos.

Todo lo anterior evidencia la importancia del desarrollo de un modelo de clasificación de células cervicales, accesible y de fácil manejo para el profesional que realiza el diagnóstico, tal que contribuya a la lectura temprana y ágil de células anormales en muestras de tejido cervical. Esto puede significar un gran impacto en regiones con difícil acceso a profesionales especializados, contribuyendo en última instancia, a la prevención y tratamiento exitoso y oportuno de esta enfermedad.

4. Objetivos

4.1. Objetivo general

Desarrollar un modelo de detección de cáncer cervical a partir de imágenes de tejido celular utilizando Máquinas de Soporte Vectorial.

4.2. Objetivos específicos

- Identificar el conjunto de características que más aporten a la clasificación de presencia de cáncer cervical.
- Proponer y evaluar una estrategia de clasificación temprana de cáncer cervical basada en técnicas de aprendizaje automático.
- Validar mediante la métrica de desempeño exactitud, la capacidad predicción del modelo de clasificación temprana de cáncer cervical.

5. Marco teórico

5.1. Cáncer cervical

El cáncer cervical surge cuando las células que recubren el cuello uterino, la parte inferior del útero, crecen sin control. Este tipo de cáncer se origina en la zona de transformación del cuello uterino, donde se encuentran dos tipos de células diferentes. Aunque el cáncer cervical a menudo se desarrolla a partir de cambios precancerosos en las células del cuello uterino, estos cambios pueden ser tratados para prevenir la aparición del cáncer. Es importante detectar el cáncer cervical en sus etapas tempranas, lo que se puede lograr a través de pruebas de detección como el Papanicolaou y la prueba de VPH. Aunque no existe tratamiento para la infección por VPH, la vacunación puede prevenirla. La mayoría de los cánceres cervicales son carcinomas de células escamosas o adenocarcinomas, y en algunos casos, pueden presentar características de ambos tipos. (American Cancer Society, 2021).

5.1.1. Tipos de células

A continuación, se define cada tipo de célula correspondiente a las clases objeto de este estudio, las cuales se pueden visualizar en la **Figura 1.**

- **Superficial – Intermediate**

  Superficial: células plenamente maduras y diferenciadas. Tienen forma poliédrica, color rojizo-anaranjado. La presencia de células superficiales depende de los niveles de estradiol (Belmonte, 2011)

  Intermediarias: son células procedentes de la descamación de la capa intermedia, de color azulado, se transforman por acción de la progesterona, por lo tanto, habrá dos tipos según la fase del ciclo ovárico: fase proliferativa y fase luteínica (Belmonte, 2011).

- **Parabasal:** Células epiteliales más pequeñas que se observan en un frotis vaginal típico y derivan de la capa basal de las células epiteliales escamosas. Son redondas o casi redondas y tienen una elevada relación núcleo/citoplasma (N:C). El citoplasma es denso y basófilo. El núcleo tiene una cromatina finamente granular y está vesiculado (Wikipedia contributors, 2022c).

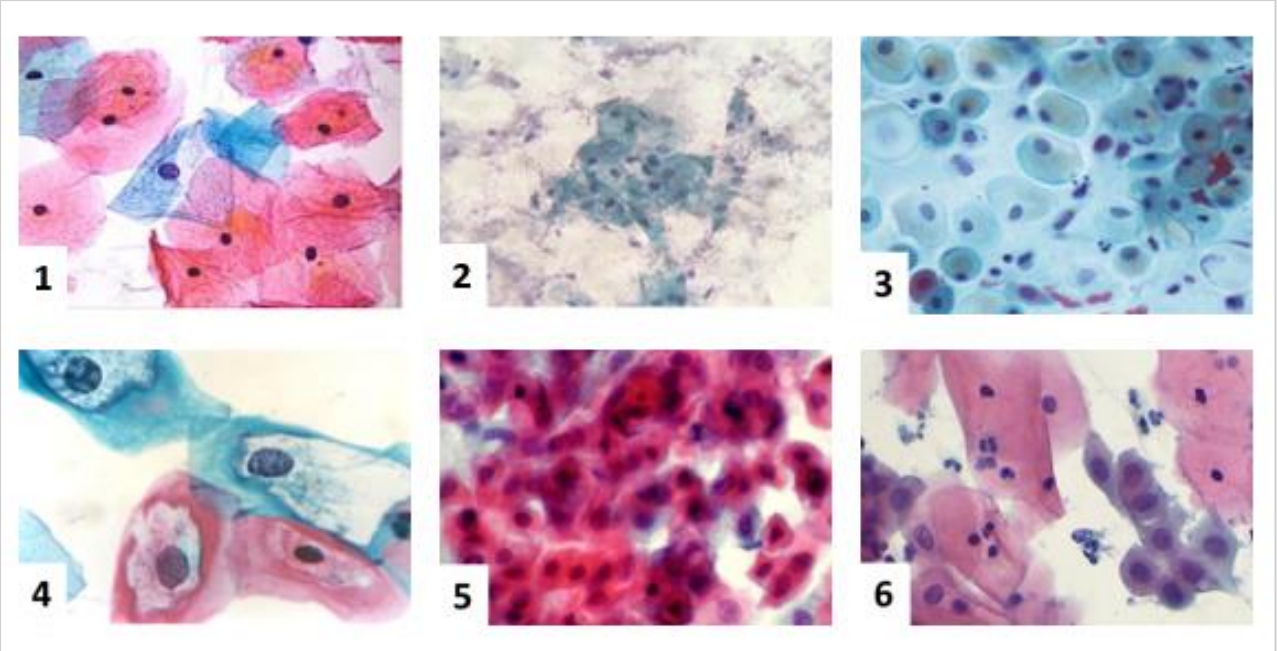

Figura 1
Tipos de células cervicales

Nota. 1. Células cervicales superficiales, 2. Células cervicales intermedias, 3. Células cervicales parabasales, 4. Células superficiales e intermedias koilocitóticas, 5. Células disqueratósicas y 6. Células metaplasicas.

- **Koilocytotic:** Célula epitelial escamosa que ha sufrido una serie de cambios estructurales, que se producen como resultado de la infección de la célula por el virus del papiloma humano (VPH). La identificación de estas células por patólogos puede ser útil para diagnosticar varias lesiones asociadas con el VPH (Wikipedia contributors, 2022b).

- **Dyskeratotic:** Este término significa que los procesos de diferenciación y queratinización están perturbados dando lugar a células queratinizadas, generalmente de pequeño tamaño y formas anormales, dentro del epitelio. Esta anormalidad es más difícil de caracterizar en la citología (núcleos agrandados con formas irregulares) (International Agency for Research on Cancer, 2004).

- **Metaplastic:** Célula diferenciada que se transforma en otro tipo de célula diferenciada. El cambio de un tipo de célula a otro puede formar parte de un proceso de maduración normal o estar causado por algún tipo de estímulo anormal. Los cambios metaplásicos suelen considerarse una fase temprana de la carcinogénesis, específicamente en el caso de las personas

con antecedentes de cáncer o que se sabe que son susceptibles de sufrir cambios carcinogénicos (Wikipedia contributors, 2022a).

En la **Figura 2** se puede visualizar el comportamiento de las células desde su estado normal hasta la transformación en un cáncer invasivo:

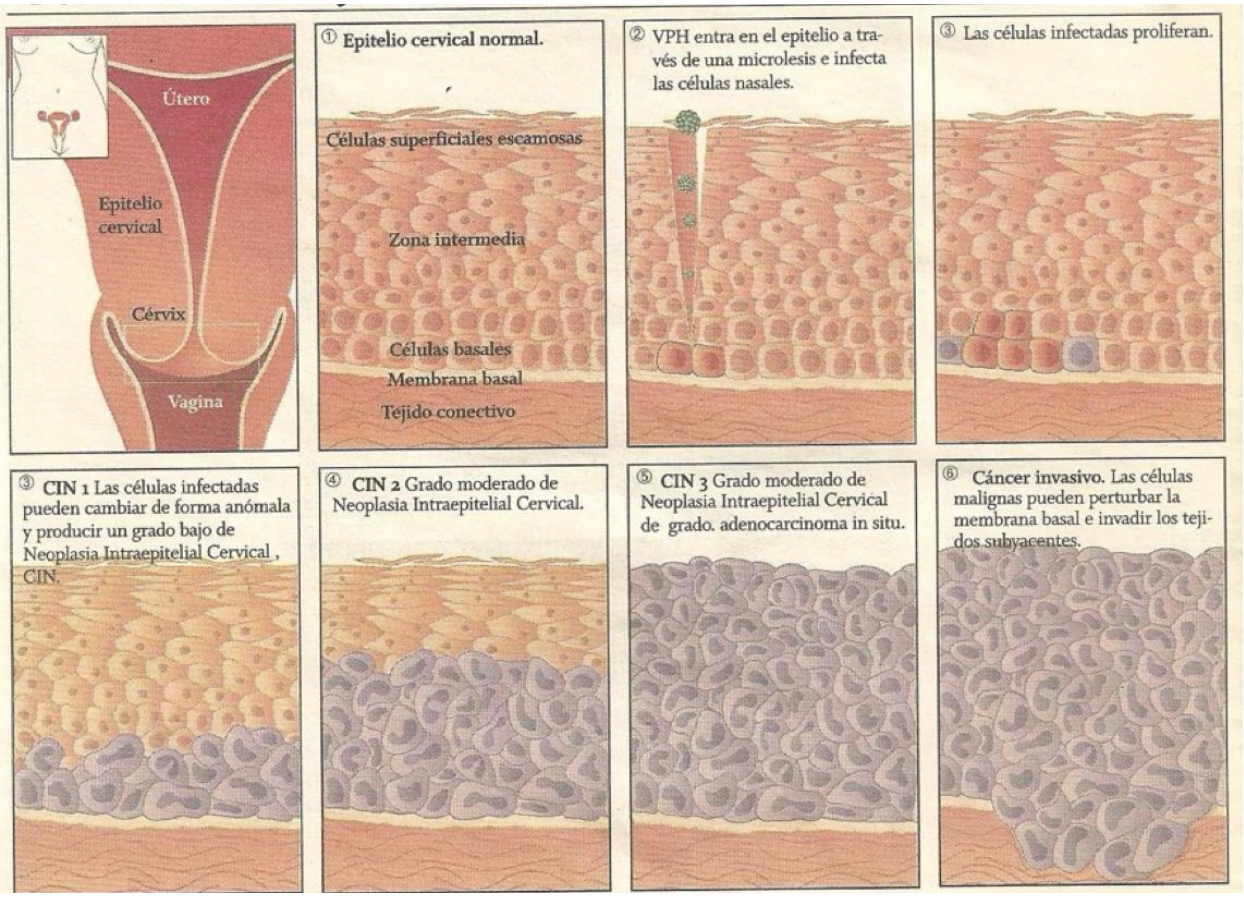

Figura 2
Desarrollo histológico del cáncer cervical

Nota. Adaptado de *Células HPV tipo 16 simuladas a ordenador,* de Asociación española de ginecología y obstetricia, consultado el 1 de mayo de 2023, (https://www.aego.es/enfermedades/cancer/cancer-de-cervix).

5.1.2. Técnicas de diagnóstico

Las pruebas para la detección del cáncer de cuello uterino son la prueba del VPH y la prueba de Papanicolaou. Estas pruebas se pueden hacer solas o al mismo tiempo (llamadas pruebas conjuntas) y se realizan durante un examen pélvico. A continuación, se da más detalle de cada una de ellas según lo encontrado en (American Cancer Society, 2021):

- **Prueba del VPH:** La infección con el virus del papiloma humano (VPH) es el factor de riesgo más importante para el desarrollo del cáncer de cuello uterino. Los médicos pueden realizar pruebas para detectar los tipos de VPH de alto riesgo que tienen más probabilidades de causar cáncer de cuello uterino, buscando fragmentos de su ADN en las células cervicales. Para realizar la prueba, se utiliza un espéculo para examinar el cuello uterino de la mujer y se toma una muestra de células utilizando un cepillo o una espátula. La muestra se coloca en una solución líquida y se envía a un laboratorio para su análisis, donde se realiza una prueba para detectar la presencia de ADN del VPH en las células. Si se encuentra ADN del VPH, esto indica que la mujer está infectada con el virus y puede tener un mayor riesgo de desarrollar cáncer de cuello uterino.

- **Prueba de Papanicolaou:** Es un procedimiento para recolectar células del cuello uterino y examinarlas en el laboratorio para detectar cáncer y precáncer. Se toma una muestra de células y moco del exocérvix, que se examina en busca de anomalías. Aunque la prueba de Papanicolaou ha sido más exitosa que cualquier otra prueba de detección para prevenir el cáncer, no es perfecta. Una de las limitaciones es que los resultados deben ser examinados por el ojo humano, lo que hace que no siempre sea posible un análisis preciso de los cientos de miles de células en cada muestra. Para abordar esto, ingenieros, científicos y médicos están trabajando juntos para mejorar la prueba. El Sistema Bethesda (TBS) es el sistema más utilizado para describir los resultados de la prueba de Papanicolaou y tiene tres categorías principales, algunas de las cuales tienen subcategorías: Negativo para lesión intraepitelial o neoplasia maligna, anomalías de las células epiteliales y otras neoplasias malignas. A pesar de que este sistema de clasificación ayuda a reducir la subjetividad en la interpretación de los resultados de la prueba, aún pueden existir algunas anomalías que se pasen por alto incluso en los mejores laboratorios.

5.2. Métrica de desempeño

La métrica de desempeño que se va a considerar para la aplicación del modelo es la exactitud (Martines, 2020):

- Verdaderos Positivos (VP): Número de datos positivos correctamente clasificados.
- Verdaderos Negativos (VN): Número de datos negativos correctamente clasificados.
- Falsos Positivos (FP): Número de datos negativos incorrectamente clasificados.
- Falsos Negativos (FN): Número de datos positivos incorrectamente clasificados.

Exactitud: esta métrica mide el porcentaje de casos que el modelo ha acertado.

$$exactitud = \frac{VP + VN}{VP + VN + FP + FN} \quad (1)$$

5.3. Técnicas de aprendizaje automático

5.3.1. Máquinas de Soporte Vectorial (SVM)

El método SVM (Support Vector Machines) se desarrolló originalmente como un clasificador binario, pero se ha expandido para incluir problemas de clasificación múltiple y regresión. Es considerado uno de los mejores clasificadores en un amplio rango de situaciones y es ampliamente utilizado en el ámbito de machine learning y aprendizaje estadístico. SVM se basa en el concepto de hiperplano de margen máximo, el cual busca maximizar la separación entre las clases o puntos de datos en un espacio de alta dimensión. Aunque el concepto de hiperplano no se generaliza de forma natural a más de dos clases, se han desarrollado varias estrategias para adaptar SVM a problemas con más de dos clases, incluyendo one-versus-one, one-versus-all y DAGSVM. A pesar de que SVM es uno de los mejores clasificadores, tiene algunas limitaciones, como su sensibilidad a la selección de parámetros y su tendencia a sobre ajustarse a conjuntos de datos pequeños (Rodrigo, 2017).

En este método se pueden utilizar diferentes hiper parámetros. A continuación, se adjuntan las definiciones de los que serán usados en este estudio, según (Rodrigo, 2017):

- **Kernel:** es una función matemática que transforma los datos de entrada a un espacio de mayor dimensionalidad. En SVM, el kernel determina el tipo de frontera de decisión que se puede crear para separar las diferentes clases en el espacio transformado. Algunos ejemplos comunes

de kernels son "linear" (lineal), "rbf" (Radial Basis Function), "poly" (polinómico) y "sigmoid" (sigmoide).

- **Gamma:** afecta la influencia de cada ejemplo de entrenamiento en la formación de la frontera de decisión. Un valor alto de gamma significa que solo los puntos cercanos tienen un impacto significativo, lo que puede llevar a fronteras de decisión más complejas y ajustadas a los datos de entrenamiento. Por otro lado, un valor bajo de gamma considera puntos más alejados en el cálculo de la frontera de decisión.

- **C:** este parámetro es el término de regularización en SVM y controla el equilibrio entre la maximización del margen y la minimización del error de clasificación. Un valor alto de C permitirá una clasificación más precisa en el conjunto de entrenamiento, pero podría llevar a un mayor riesgo de sobreajuste. Por el contrario, un valor bajo de C dará mayor importancia al margen amplio, lo que puede generar una clasificación menos precisa, pero con una mejor generalización.

- **Degree:** solo es relevante cuando se utiliza el kernel polinómico. Representa el grado del polinomio utilizado en la función de kernel. Un valor más alto de degree dará lugar a fronteras de decisión más complejas y no lineales.

- **Coef0:** también se utiliza en los kernels polinómico y sigmoidal. Define el término independiente en estas funciones de kernel y controla la influencia de los términos de orden superior. Un valor alto de coef0 puede llevar a fronteras de decisión más flexibles.

5.3.2. t-SNE (del ingles T-distributed Stochastic Neighbor Embedding)

Es una herramienta utilizada para visualizar datos de alta dimensión. Funciona mediante la conversión de las similitudes entre puntos de datos en probabilidades conjuntas y trata de minimizar la divergencia de Kullback-Leibler entre las probabilidades conjuntas de la incrustación de baja dimensión y los datos de alta dimensión. Es importante tener en cuenta que la función de coste de t-SNE no es convexa, lo que significa que, con diferentes inicializaciones, podemos obtener resultados diferentes (scikit-learn 1.2.2 documentation, 2021).

5.4. Técnicas de validación

5.4.1. Técnica de selección de características

Para la elaboración del modelo, lo primero que se realiza es la extracción de características RGB sobre las imágenes, lo cual consiste en identificar, analizar y representar información de color en base a los valores de rojo, verde y azul en cada píxel de la imagen. En particular, la utilización de características RGB permite medir la intensidad de cada canal de color y obtener información relevante sobre la composición cromática de las células en la muestra.

Las características RGB sobre las imágenes usadas para el análisis fueron: la intensidad promedio, la cual es una medida básica de la cantidad de luz o brillo presente en la imagen, la suavidad, que se refiere a la cantidad de cambios de intensidad en la imagen, la uniformidad que se refiere a la distribución de intensidades en la imagen, la entropía, la cual se utiliza para medir la complejidad de la imagen y se refiere a la cantidad de información o desorden presente en la imagen (Pérez & Valente, 2018). Así mismo, el tercer momento, el cual es una medida estadística que se utiliza en análisis de imágenes para describir la forma y la textura de una imagen. El tercer momento mide la asimetría de la distribución de intensidad de los píxeles en una imagen. Una distribución perfectamente simétrica tendría un tercer momento igual a cero (Gonzalez & Woods, 2008).

Por último, se usó en escala de grises el contraste promedio, que hace referencia a una de las más importantes características que se pueden medir en una imagen, dado que se refiere a la relación entre la intensidad más alta y la más baja en una imagen, es decir a la diferencia de luminancia entre diferentes áreas de la imagen (Gonzalez & Woods, 2008). Una imagen con alto contraste tiene áreas con luminancias muy diferentes, mientras que una imagen con bajo contraste tiene áreas con luminancias similares. El contraste es una característica importante porque puede afectar la capacidad de una persona o una máquina para detectar objetos o características en una imagen.

5.4.2. Técnica de eliminación de datos atípicos

Para garantizar la calidad de los datos y evitar errores en los análisis de los modelos, es necesario realizar la detección y extracción de datos atípicos. Una técnica ampliamente utilizada para este propósito es el método LOF (Local Outlier Factor), que se basa en la identificación de

puntos de baja densidad en el espacio de características y su comparación con los vecinos cercanos para determinar si son atípicos o no (Breunig et al., 2000). La aplicación de esta técnica puede contribuir a la mejora de la precisión y robustez del modelo.

5.4.3. Técnicas de escalamiento

En este análisis se evaluarán 3 métodos de escalamiento: Robusto, Estándar y Min-Max. A continuación, se adjuntan las definiciones de cada uno según (Machine Learning Geek, 2020):

- **Escalamiento robusto:** es una técnica de escalado que se destaca por su capacidad para tratar con valores atípicos en los datos. Por defecto, utiliza el rango intercuartil (IQR), que se define como la diferencia entre el tercer y primer cuartil de los datos. Además, se puede ajustar manualmente el rango cuantílico mediante el parámetro quantile_range. Al aplicar el Robust Scaler, tanto la mediana como la escala de los datos se ajustan en función del rango cuantílico especificado, permitiendo una mejor visualización y análisis de los datos.
- **Escalamiento estándar:** el escalado estándar se basa en la suposición de que los datos dentro de cada característica siguen una distribución normal estándar, lo que implica que la distribución de los datos se centra en 0 y se desvían con una desviación estándar de 1, eliminando la media. Esto permite que las características se encuentren en la misma escala y sean comparables entre sí, mejorando el rendimiento de los modelos de aprendizaje automático. Además, el escalado estándar reduce la influencia de los valores extremos presentes en los datos, lo que puede mejorar aún más el rendimiento de los modelos.
- **Escalamiento Min-Max:** el Min-Max Scaler es un método de escalado que ajusta las características de un conjunto de datos dentro de un rango específico, comúnmente [0,1] o [-1,1] en caso de valores negativos. Es una buena opción cuando la desviación estándar es pequeña o los datos no siguen una distribución gaussiana. Este algoritmo asegura que las características estén dentro del rango definido, lo que mejora el rendimiento de los modelos de aprendizaje automático.

5.4.4. Técnicas de reducción de dimensión

- **Independent Component Analysis (ICA):** el Análisis de Componentes Independientes (ICA, por sus siglas en inglés) es una técnica de procesamiento de señales que tiene como objetivo separar una señal multivariada en componentes independientes. La idea es descomponer señales complejas en diferentes fuentes originales para su análisis individual. Esta técnica se basa en la suposición de que cada una de las señales originales no está correlacionada con las demás, lo que permite identificar y extraer patrones ocultos en los datos. ICA tiene aplicaciones en diversas áreas, como la eliminación de ruido de las señales, la separación de fuentes de audio y video, y el análisis de datos de neuroimagen, entre otros. (Jouan-Rimbaud Bouveresse & Rutledge, 2016).
- **Principal Component Analysis (PCA):** es una técnica utilizada para reducir la dimensionalidad de un conjunto de variables correlacionadas al transformarlas en un conjunto de variables no correlacionadas, conocidas como componentes principales. Esta técnica se utiliza comúnmente para reducir la redundancia y complejidad de los datos (Johnson et al., 2007).
- **Linear Discriminant Analysis (LDA):** el Análisis Discriminante Lineal (LDA, por sus siglas en inglés) es una técnica popular de reducción de dimensionalidad utilizada en problemas de clasificación supervisada en aprendizaje automático. Se utiliza comúnmente como un paso de preprocesamiento para modelar las diferencias entre clases en aplicaciones de clasificación de patrones. Esta técnica es particularmente útil cuando se requiere la separación eficiente de dos o más clases que tienen múltiples características. En situaciones donde las clases se superponen al clasificarlas utilizando solo una característica, el aumento del número de características puede ayudar a resolver el problema de superposición en el proceso de clasificación. El LDA es una técnica eficaz para abordar este problema y lograr una separación óptima de las clases con múltiples características (Schlagenhauf, 2022).

6. Estado del arte

Desde los últimos 5 años se han tenido diversos estudios relacionados con el tema, todos enfocados en alcanzar, a través de diferentes modelos de clasificación y técnicas aprendizaje, una detección temprana del cáncer de cuello uterino. Se han usado desde máquinas de soporte vectorial hasta redes neuronales.

Uno de los primeros estudios realizados y que logró su objetivo de ser un referente para futuras técnicas de clasificación, es el realizado en 2018 por Marina E. Plissiti et al. (Plissiti et al., 2018), en el cual presentan una base de datos de imágenes de frotis de Papanicolaou, en la que las células se clasifican en cinco clases diferentes, en función de sus características citomorfológicas. Los autores proponen modelos basados en máquinas de soporte vectorial y redes neuronales profundas para clasificar imágenes de cuello uterino a partir de sus características de celda, píxeles y citomorfológicas. El modelo basado en redes convolucionales con la característica de color alcanzó una exactitud del 95.0% y para el modelo basado en máquinas de soporte vectorial en la característica de profundidad logró 94.4% de exactitud y con características celulares un 91.6%. Este estudio se basa en el presente artículo, el cual sienta bases sólidas basadas en técnicas como la extracción de características y la aplicación de SVM. El objetivo es incorporar otras técnicas de análisis, como el escalamiento de datos y la reducción de dimensionalidad, con el fin de mejorar la exactitud del modelo.

En el año 2020, los autores Basak et al. (Basak et al., 2021), llevaron a cabo un estudio en el campo de la clasificación de imágenes cervicales utilizando técnicas de aprendizaje profundo, los autores lograron extraer características de imagen significativas mediante el análisis de componentes principales (PCA) y la optimización del lobo gris (GWO). Como resultado, el modelo propuesto logró una precisión de clasificación de imágenes del 98.3% para la base de datos de Herlev y 97,8% en la base de datos de SIPaKMeD.

Para el año 2021, en el estudio de Manna et al. (Manna et al., 2021), utilizaron redes neuronales convolucionales con un enfoque basado en rangos difusos para la extracción de características, logrando una precisión del 95,4% y una sensibilidad del 98,5% en su entorno de 5

clases, en la base de datos de SIPaKMeD. Por otro lado, en el estudio de Mohammed A et al. (Mohammed A et al., 2021), los autores evaluaron la base de datos SIPaKMeD mediante el uso de arquitecturas de redes neuronales profundas pre-entrenadas y métricas de rendimiento como precisión, recall y F1-score. Con DenseNet169, obtuvieron un rendimiento promedio de 99.0%, 97.4%, 97.4% y 97.4%, respectivamente, superando la precisión de referencia propuesta por los creadores del conjunto de datos en un 3.7%.

En 2022, se presentan dos estudios relevantes de los autores Pramanik et al. (Pramanik et al., 2022) y Fekri-Ershad S & Fadhil M (Fekri-Ershad S & Fadhil M, 2023). El primer estudio aplicó modelos de aprendizaje de transferencia en la base de datos SIPaKMeD, y logró rendimientos de 95.3%, 93.9% y 96.4% utilizando Inception V3, MobileNet V2 e Inception ResNet V2, respectivamente. Por otro lado, en el segundo estudio los autores emplearon redes neuronales de perceptrón multicapa (MLP) en la base de datos Herlev, obteniendo una precisión del 99.2% para la clasificación binaria y del 97.6% para la clasificación multiclase.

Con lo anterior, se puede afirmar que la aplicación de tecnologías de aprendizaje automático es efectiva en la detección no solo del cáncer cervical, sino también en otras tareas de clasificación de imágenes médicas, lo que brinda una valiosa herramienta para los profesionales de la salud.

7. Metodología

El flujo de trabajo que se pretende realizar, el cual se puede visualizar en la **Figura 3**, comienza con el preprocesamiento y preparación de la data, donde se limpia y se transforma la data para que sea adecuada para el modelo. Luego, se aplica el modelo de Máquinas de Soporte Vectorial en la fase de entrenamiento y prueba, donde se entrenará el modelo y posterior se evaluará con los datos de prueba. A continuación, se ajustarán los hiper parámetros para realizar diferentes experimentos y se evaluarán los resultados en términos de la exactitud. Finalmente, con el análisis de estas métricas, se llegará a conclusiones sobre qué ajustes de hiper parámetros son los más adecuados para la tarea en cuestión.

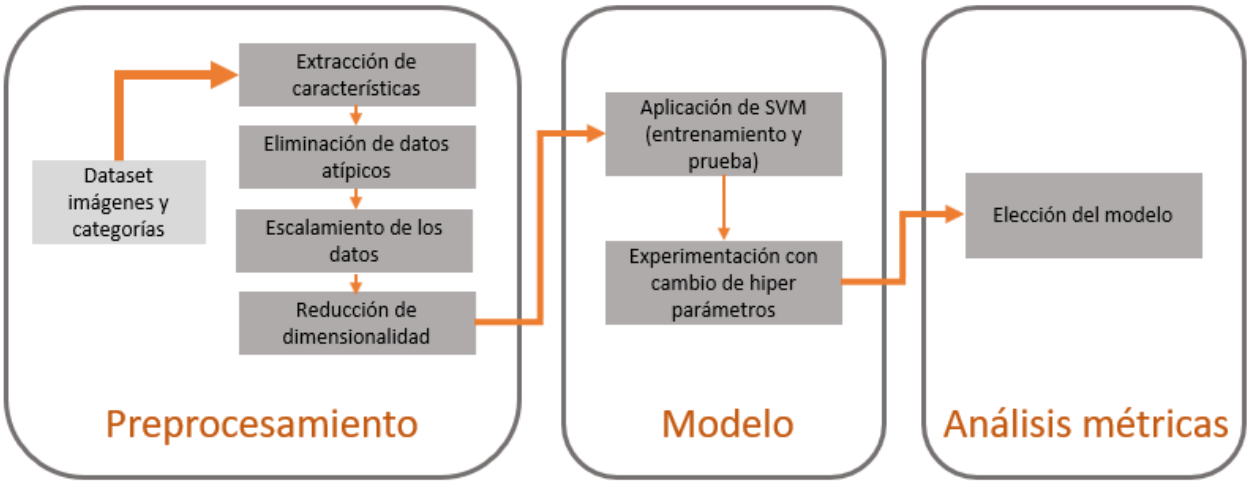

Figura 3
Flujo de trabajo del proyecto

7.1. Fase de preprocesamiento de la data

A partir del entendimiento general del problema y los datos, el primer paso y uno de los más relevantes de cara a la aplicación del modelo, es el preprocesamiento de la data, en este caso de las 4.049 imágenes de células contenidas en la base de datos SIPaKMeD, las cuales fueron recortadas y divididas en 5 grandes categorías, enumeradas en la **Tabla 1**, y que para los análisis posteriores se denotarán de la siguiente manera: K: Koilocytotic; S: Superficial – Intermediate; M: Metaplastic; D: Dyskeratotic y P: Parabasal.

7.1.1. Extracción de características

Para preparar la data, se realizó un proceso de extracción de características de imágenes en cada una de las dimensiones RGB de todas las imágenes recortadas dentro las 5 categorías. En este caso, se han extraído siete características diferentes de cada imagen en el conjunto de datos. Estas características incluyen el promedio de intensidad de cada canal de color, la suavidad de la imagen, la uniformidad, el tercer momento, la entropía, la desviación estándar y la mediana. Así mismo, se evalúa una característica adicional, pero medida bajo la escala de grises: El contraste promedio o luminancia.

El proceso de extracción de características lee la información de las diferentes categorías (K, S, M, D, P), y por cada una de las categorías se recolecta la clase a la que pertenece, la ubicación de la imagen analizada y las características previamente mencionadas, las cuales se toman de cada una de las imágenes.

7.1.2. Eliminación de datos atípicos

Considerando que se pueden tener imágenes con un comportamiento anómalo, según los datos obtenidos con los vectores de características, y en aras de garantizar la calidad de los datos y evitar errores en el análisis del modelo, se realiza la detección y extracción de datos atípicos a través del método LOF (Local Outlier Factor).

Para visualizar los datos atípicos de las 5 clases, se utiliza el método t-SNE, ya que la data se encontraba en múltiples dimensiones. Para representar esta información en un gráfico de dos dimensiones, se empleó la técnica de Análisis de Componentes Principales (PCA) y se seleccionan las componentes con mayor varianza del análisis para su representación gráfica. Así mismo, se analizan las imágenes de los datos atípicos para cada una de las clases, con el fin de observar su comportamiento y características visuales.

7.1.3. Escalamiento de los datos

Antes de aplicar el modelo, es importante evaluar el comportamiento de los datos mediante diferentes métodos de escalamiento, ya que algunas características pueden tener un rango de valores mucho mayor que otras, lo que podría afectar negativamente el rendimiento de este. Escalar los datos para los algoritmos de aprendizaje automático puede mejorar significativamente el

rendimiento del modelo, ya que ayuda a igualar el peso de las características y a mejorar la precisión de la distancia euclidiana. Para ello, se procede a crear conjuntos de entrenamiento y prueba utilizando la función train_test_split de scikit-learn. Se utiliza un tamaño de prueba del 20% del conjunto de datos.

Se instancian los tres métodos elegidos: Robusto, Estándar y Min-Max. Los conjuntos de datos escalados y sin escalar son utilizados posteriormente para aplicar las técnicas de reducción de dimensionalidad, que en conjunto permitan comparar los resultados de los diferentes métodos para determinar cuál de ellos tiene un mejor desempeño en el modelo en términos de la métrica exactitud.

7.1.4. Reducción de dimensionalidad

Se utilizan tres técnicas de reducción de dimensionalidad: Análisis de Componentes Independientes (ICA), Análisis de Componentes Principales (PCA) y Análisis Discriminante Lineal (LDA). Estas técnicas permiten simplificar la representación de datos complejos y de alta dimensión, lo que facilita su análisis y visualización, y ayudan a evitar el sobreajuste.

Para la aplicación de las técnicas se usa el criterio de la varianza, el cual permite encontrar el número de componentes óptimo. Lo anterior dado que, si se mantienen demasiadas componentes, se puede incluir ruido y reducir la capacidad del modelo para generalizar a nuevos datos. Por otro lado, si se mantienen muy pocas componentes, se puede perder información importante.

En primer lugar, se utiliza la técnica de PCA a través de una función llamada find_optimal_pca_components que ajusta el modelo PCA sobre los datos y determina el número de componentes principales necesarios para explicar al menos el 90% de la varianza total. El proceso se hace sobre cada una de las representaciones de los datos con diferentes escaladores y para cada representación de datos se calcula la cantidad óptima de componentes principales. Luego, se aplica PCA con el número óptimo de componentes, se guarda la transformación de los datos y se almacena el modelo PCA y los datos transformados en un diccionario llamado feature_hash para su posterior uso.

A continuación, se aplica la técnica de ICA mediante una función llamada find_optimal_components que utiliza el algoritmo FastICA para encontrar los componentes independientes. Se calcula la cantidad de componentes necesarios para explicar al menos el 90% de la varianza en las componentes independientes resultantes. Luego, se aplica ICA con esa cantidad de componentes a cada conjunto de datos con diferentes escaladores y se guarda la transformación ICA resultante en una estructura de datos.

Por último, se utiliza la técnica de LDA para cada conjunto de datos en diferentes escalas, se utiliza la función get_optimal_lda_components para determinar el número de componentes necesarios para representar al menos el 95% de la varianza original de los datos. Luego, se realiza la reducción de dimensionalidad y se almacena el resultado en feature_hash.

Una vez obtenidas las representaciones de los datos con los diferentes escaladores y algoritmos, se procede a verificar cuál es la mejor representación de datos basados en la clasificación obtenida por una máquina de soporte vectorial.

En la **Figura 4** se puede tener una visual de cómo sería el comportamiento de esta fase de preprocesamiento de la data:

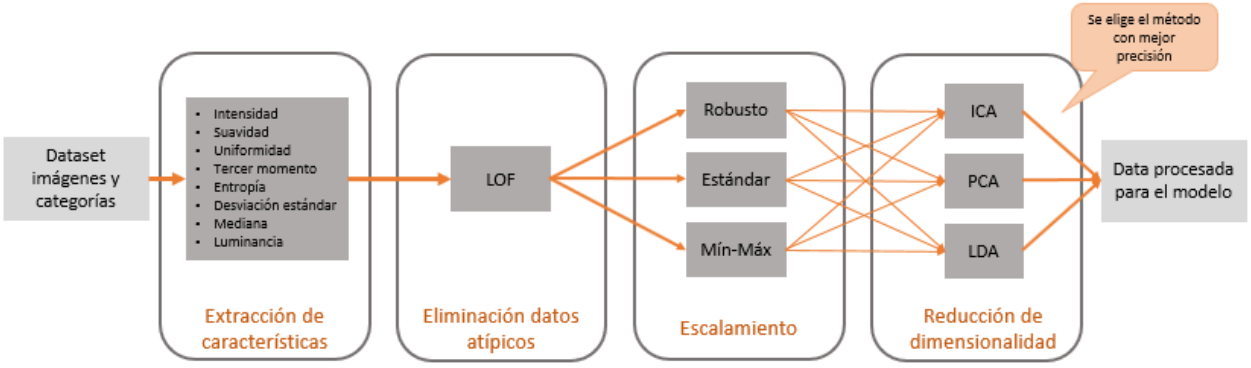

Figura 4
Etapa metodológica relacionada con la fase de preprocesamiento de las imágenes de células cervicales

7.2. Desarrollo del modelo

7.2.1. Aplicación de SVM (entrenamiento y prueba)

Después de analizar la exactitud de los métodos de escalamiento y reducción de dimensionalidad, se puede aplicar el método de Máquinas de Soporte Vectorial a los datos preprocesados con el mejor método de escalamiento y reducción de dimensionalidad. El objetivo de SVM es encontrar el hiperplano que mejor separa las diferentes clases en el conjunto de datos. Para ello, se divide el conjunto de datos en entrenamiento y prueba, utilizando la proporción 80-20. El modelo se evalúa bajo la exactitud definida en la ecuación (1).

7.2.2. Experimentación con cambio de hiper parámetros

Para mejorar el rendimiento del modelo, se ajustan los parámetros, como el tipo de kernel, la regularización C, el valor de gamma y el valor constante en la función de decisión (coef0). La idea es probar diferentes combinaciones para encontrar las que generen el mejor rendimiento en términos de exactitud al clasificar los datos de prueba. Para cada combinación, se crea un modelo SVM utilizando los valores correspondientes de los hiper parámetros y se entrena con los datos de entrenamiento.

Una vez que se ajusta el modelo SVM, se evalúa su exactitud en el conjunto de prueba mediante la ecuación (1). Se selecciona el modelo que logra la exactitud deseada (95.0%), el cual puede ser usado para hacer predicciones en nuevos conjuntos de datos.

8. Resultados y Análisis

8.1. Preprocesamiento de la data

Para el análisis a realizar se tienen datos clínicos de un dataset de imágenes clasificados en 5 grupos (estos últimos clasificados a su vez en 3 grupos: normal, anormal y benigno). Los datos provienen de una base de datos open source llamada SipakMed (SIPaKMeD, 2020). La base de datos contiene 4,049 imágenes de células que se han etiquetado manualmente a partir de 966 imágenes de muestra de tejido cervical de resolución [1536 px, 2048 px]. Las imágenes se adquirieron a través de una cámara CCD (Infinity 1 Lumenera) adaptada a un microscopio óptico (OLYMPUS BX53F) (Basak et al., 2021). En la **Tabla 1** se ilustra la distribución de clases correspondiente a las etiquetas asignadas a cada una de las células etiquetadas. Dado que las imágenes categorizadas provienen de recortes de las imágenes originales, las resoluciones de estas son diversas.

Tabla 1
Tipos de células de tejido cervical

Tipo de Célula	Categoría	Número Imágenes de células
S: Superficial - Intermediate	Normal	831
P: Parabasal	Normal	787
K: Koilocytotic	Anormal	825
D: Dyskeratotic	Anormal	813
M: Metaplastic	Benigna	793

En la extracción de características, con las 7 características de imagen en cada uno de los canales RGB y la característica correspondiente a la luminancia, se logran obtener 21 valores en total por imagen. En la **Tabla 2** se muestra un ejemplo de un registro, en este caso dentro de la categoría S.

Tabla 2

Ejemplo extracción características en una imagen de la categoría S

R	Intensidad promedio	feature_0	183.323388
	Suavidad	feature_1	0.01802
	Uniformidad	feature_2	0.028836
	Tercer Momento	feature_3	-917.959535
	Entropía	feature_4	6.254622
	Desviación estándar	feature_5	20.068231
	Mediana	feature_6	184.0
G	Intensidad promedio	feature_7	139.200186
	Suavidad	feature_8	0.017287
	Uniformidad	feature_9	0.035768
	Tercer Momento	feature_10	328.495952
	Entropía	feature_11	6.395571
	Desviación estándar	feature_12	23.040231
	Mediana	feature_13	139.0
B	Intensidad promedio	feature_14	72.627459
	Suavidad	feature_15	0.013979
	Uniformidad	feature_16	0.021424
	Tercer Momento	feature_17	559.514412
	Entropía	feature_18	6.293955
	Desviación estándar	feature_19	22.126087
	Mediana	feature_20	67.0
	Luminancia	feature_21	124.324153

En el análisis de los datos atípicos, con el método LOF se calculó para cada punto de datos y se comparó con el LOF de sus vecinos cercanos, en este caso 5 vecinos cercanos, con una contaminación de 5.0%. Estos hiper parámetros fueron obtenidos bajo un análisis del MSE (error cuadrático medio), dentro de cada categoría con los siguientes rangos de valores:

- neighbors_values = [5, 10, 15, 20]
- contamination_values = [5, 50, 5]

A continuación, se detallan en la **Tabla 3** los datos atípicos encontrados dentro de cada una de las categorías:

Tabla 3
Cantidad de datos atípicos por categoría

Categoría	Datos atípicos	Datos totales	Porcentaje de atípicos/total (%)
K	42	825	5.0%
D	41	813	5.0%
M	40	793	5.0%
P	40	787	5.0%
S	42	831	5.0%

Una vez obtenidos los datos atípicos, se visualiza su distribución haciendo uso de la técnica t-SNE. En la **Figura 5**, se aprecia como los datos atípicos (en morado) se ubican en los extremos de la distribución de datos.

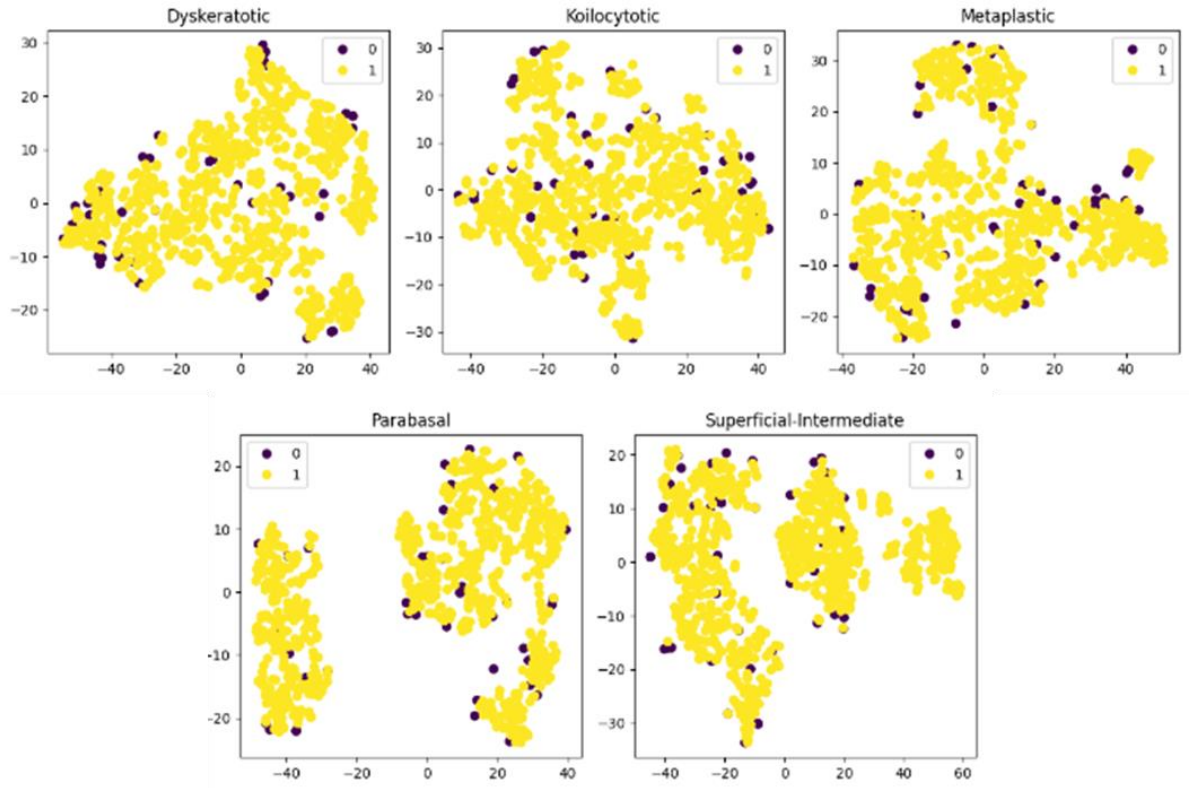

Figura 5
Scatterplot de los datos. Datos atípicos: círculos morados.

Sin embargo, dado que la naturaleza de la base de datos son imágenes, en la **Figura 6** se realiza una visualización por categoría de algunas de las imágenes atípicas, como estrategia de confirmación visual de características diferentes a las imágenes de buena calidad.

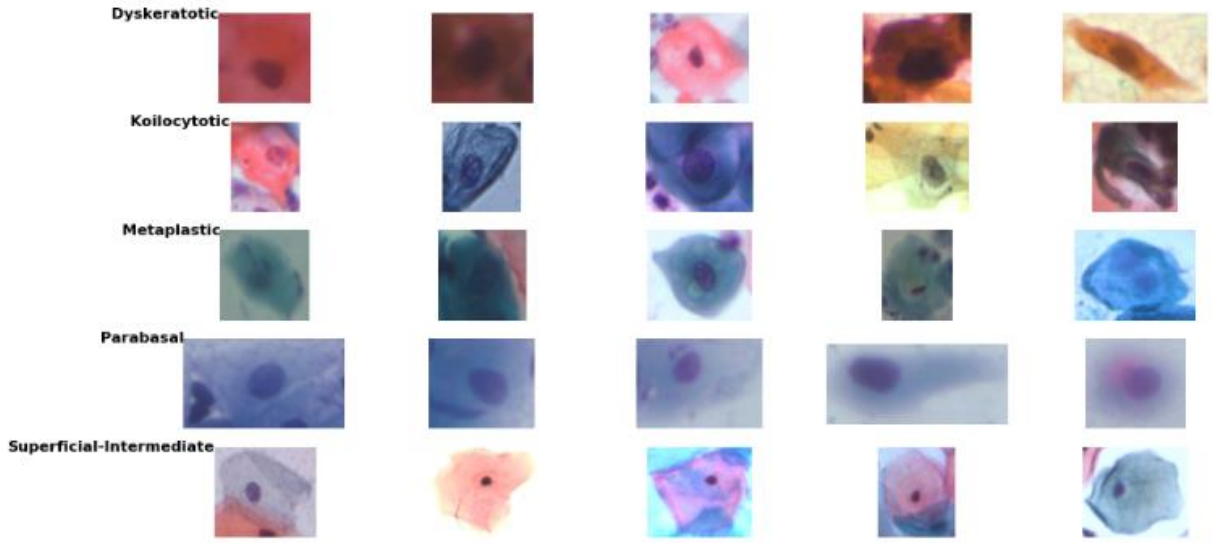

Figura 6
Ejemplos imágenes correspondientes a algunos de los datos atípicos por categoría

8.2. Desarrollo del modelo

Luego de tener la data sin datos atípicos, se aplican los modelos de reducción de dimensionalidad PCA, ICA y LDA, con los diferentes escaladores Estándar, Robusto, Min-Max y sin escalar. Se obtiene para PCA que el número de componentes óptimo en cada escalador es: 7, 7, 5 y 2, respectivamente. Para ICA el número óptimo de componentes en todos los escaladores es 20 y para LDA es 4.

Con lo anterior, se evalúa la mejor representación de datos basados en la clasificación obtenida por una máquina de soporte vectorial y se identifica en la **Tabla 4** que los algoritmos ICA y LDA son los que presentan mejor exactitud en el entrenamiento, mientras que ICA es el que mejor exactitud presenta en la prueba cuando los datos no son escalados.

Para una mejor comprensión de los resultados, se presenta en **Figura 7** los gráficos t-SNE tanto de los datos originales como de los datos reducidos en su dimensionalidad. La visualización

de los datos originales en t-SNE por clase, permite observar que no hay una buena agrupación de los datos y la representación de los datos sin escalar utilizando ICA, permite identificar una mejor agrupación de estos.

Tabla 4

Exactitud diferentes escaladores y métodos de reducción de dimensionalidad.

Escalador	Método de reducción	Exactitud train	Exactitud test
features_rob_scaled	ica	95.0%	20.0%
features_std_scaled	ica	95.0%	20.0%
features_mm_scaled	ica	94.9%	20.0%
features_no_scaled	ica	94.8%	93.8%
features_no_scaled	lda	90.4%	88.9%
features_std_scaled	lda	90.4%	20.5%
features_rob_scaled	lda	90.4%	20.5%
features_mm_scaled	lda	90.4%	20.5%
features_std_scaled	pca	89.3%	20.0%
features_mm_scaled	pca	88.2%	20.5%
features_rob_scaled	pca	88.0%	20.0%
features_no_scaled	pca	55.8%	56.0%

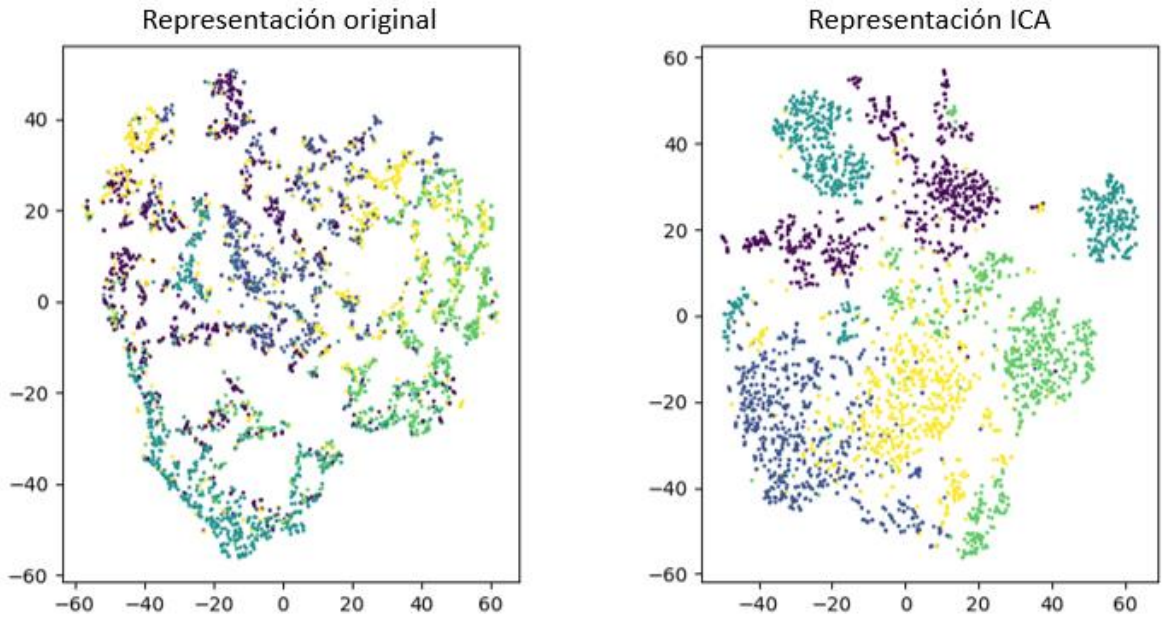

Figura 7
Visualización de agrupación de datos por t-SNE

Nota. Amarillo: categoría K, Azul: categoría D, Cian: categoría M, Verde: categoría S, Morado: categoría P.

Una vez se dedujo que el método de reducción de dimensionalidad que representa mejor los datos y maximiza su separación es el de Análisis de Componentes Independientes (ICA). Se procede con esta información a aplicar este método en conjunto con el modelo de Máquinas de Vectores de Soporte (SVM).

Para ello, se utiliza el módulo SVC de la biblioteca sklearn. Con el fin de encontrar el mejor clasificador, se emplea una grilla de hiper parámetros donde se evalúa la exactitud y se selecciona el clasificador que maximiza esta métrica. Esta estrategia permite identificar el modelo de SVM óptimo para los datos dados, al ajustar y seleccionar los hiper parámetros adecuados. De esta manera, se busca obtener el mejor rendimiento y desempeño en la tarea de clasificación. La grilla de hiper parámetros utilizada fue:

- C: 0.001, 0.01, 0.1, 1, 10, 100
- gamma: scale, auto, 0.001, 0.01, 0.1, 1
- kernel: rbf, sigmoid, linear, poly

- degree: 2, 3, 4
- coef0: -1, 0, 1

Al analizar esta grilla de hiper parámetros, se observa que el *kernel* más efectivo es el "rbf", la mejor opción para el parámetro *gamma* es "scale", el valor óptimo para *C* es 10 y los parámetros *degree* y *coef0* deben quedar en 2 y -1 respectivamente. Con lo anterior se logra una exactitud en el modelo de 95.5%.

Para la extracción de características, se tomó como base el artículo realizado en 2018 por Marina E. Plissiti et al. (Plissiti et al., 2018), en este artículo se reportó una exactitud del 91.6% utilizando características celulares. Sin embargo, con el enfoque de este estudio, se obtuvo una exactitud de 95.5%, utilizando una reducción de dimensionalidad con ICA y un SVM con los parámetros: *kernel*=rbf, *gamma*=scale, *C*=10, *degree*=2 y *coef0*=-1.

8.3. Estructura del repositorio

El repositorio asociado al detalle del desarrollo del modelo se encuentra disponible en: https://github.com/Alberto-San/ExperimentosMonografia/tree/main

Los notebooks están organizados en el siguiente orden:

- **Extracción de Características.ipynb:** contiene el código para extraer las características relevantes de los datos de las imágenes de la base de datos de SIPaKMeD.
- **Análisis de Distribución de Datos.ipynb:** en este notebook se analizan las distribuciones de los datos antes y después de aplicar la técnica de extracción de características. Se explora el método de Local Outlier Factor para filtrar los datos atípicos.
- **Preprocesamiento.ipynb:** este notebook aloja todo el preprocesamiento de los datos. Se encuentra la mejor representación de los datos que maximiza su separación empleando diferentes métodos de escalamiento y de reducción de dimensionalidad.
- **Clasificación & Finne Tunning.ipynb:** en este notebook se realiza la selección de la mejor máquina de soporte vectorial usando una grilla de hiper parámetros, y se escoge aquella que maximiza la precisión (accuracy).

Se recomienda seguir el orden en el que se presentan los notebooks para una mejor comprensión del proceso completo. Cada notebook contiene comentarios detallados que explican el código y los resultados obtenidos.

9. Conclusiones

- El modelo propuesto en este trabajo, basado en Máquinas de Soporte Vectorial (SVM), ha demostrado una mayor exactitud del 3.9% en comparación con el enfoque presentado por Marina E. Plissiti et al. (Plissiti et al., 2018). Los resultados obtenidos fueron de un 95.5% de exactitud frente al 91.6% del trabajo previo. Esto fue posible mediante la configuración de los parámetros kernel=rbf, gamma=scale, C=10, degree=2 y coef0=-1. Estos resultados resaltan la efectividad de esta propuesta como una alternativa altamente eficaz para respaldar el diagnóstico temprano y preciso del cáncer cervical.
- En este estudio, se utilizó el método LOF para identificar los datos atípicos en el conjunto de imágenes de células cervicales. Se encontró que cada categoría de células tenía aproximadamente el 5% de datos atípicos. Al visualizar la distribución de estos datos mediante t-SNE, se observó que se encontraban en los extremos de la distribución general de los datos, relacionados con imágenes de baja calidad.
- En relación con la reducción de dimensionalidad, se exploraron diferentes técnicas, como PCA, ICA y LDA, en combinación con distintos escaladores. Se determinó que la técnica de Análisis de Componentes Independientes (ICA) fue la más efectiva para representar los datos y maximizar su separación.

Referencias

American Cancer Society. (2021). *Cervical Cancer*. American Cancer Society. https://www.cancer.org/cancer/cervical-cancer.html

Asociación Española de Ginecología y Obstetricia. (n.d.). *Cáncer de Cervix*. Asociación Española de Ginecología y Obstetricia. Retrieved April 30, 2023, from https://www.aego.es/enfermedades/cancer/cancer-de-cervix

Basak, H., Kundu, R., Chakraborty, S., & Das, N. (2021, July 7). *Cervical Cytology Classification Using PCA and GWO Enhanced Deep Features Selection*. SN Computer Science. https://link.springer.com/article/10.1007/s42979-021-00741-2

Belmonte, L. (2011). *Citología en ginecología*. https://www.chospab.es/web/area_medica/obstetriciaginecologia/docencia/seminarios/2011-2012/sesion20110615_1.pdf

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000, June). *LOF: identifying density-based local outliers*. ACM SIGMOD International Conference on Management of Data. https://dl.acm.org/doi/10.1145/335191.335388

DANE. (2022, October 18). *Situación de las mujeres rurales desde las estadísticas oficiales*. Dane.Gov.Co. https://www.dane.gov.co/files/investigaciones/notas-estadisticas/oct-2022-nota-estadistica-mujer-rural-presentacion.pdf

Fekri-Ershad S, & Fadhil M. (2023, February 12). *Developing a Tuned Three-Layer Perceptron Fed with Trained Deep Convolutional Neural Networks for Cervical Cancer Diagnosis*. MDPI Open Acces Journals.

Gonzalez, R. C., & Woods, R. E. (2008). *Digital Image Processing Second Edition*. Prentice Hall. https://www.academia.edu/25992484/Digital_Image_Processing_Second_Edition

International Agency for Research on Cancer. (2004). *Histopathology of the uterine cervix - digital atlas*. IARC CancerBase No. 8. Histopathology and cytopathology of the uterine cervix - digital atlas - glossary (iarc.fr)

Johnson, G. W., Ehrlich, R., Full, W., & Ramos, S. (2007). Principal components analysis and receptor models in environmental forensics. *Introduction to Environmental Forensics*, 207–272. https://doi.org/10.1016/B978-012369522-2/50008-7

Jouan-Rimbaud Bouveresse, D., & Rutledge, D. N. (2016). Independent Components Analysis: Theory and Applications. *Data Handling in Science and Technology*, *30*, 225–277. https://doi.org/10.1016/B978-0-444-63638-6.00007-3

Machine Learning Geek. (2020, November 5). *Feature Scaling: MinMax, Standard and Robust Scaler*. Machine Learning Geek. https://machinelearninggeek.com/feature-scaling-minmax-standard-and-robust-scaler/

Manna, A., Kundu, R., Kaplun, D., Sinitca, A., & Sarkar, R. (2021, July 15). *A fuzzy rank-based ensemble of CNN models for classification of cervical cytology*. Scientific Reports. https://www.nature.com/articles/s41598-021-93783-8

Martines, J. (2020, October 9). *Precision, Recall, F1, Accuracy en clasificación*. IArtificial.Net. https://www.iartificial.net/precision-recall-f1-accuracy-en-clasificacion/#:~:text=Cuando%20necesitamos%20evaluar%20el%20rendimiento,utilidad%20pr%C3%A1ctica%20con%20un%20ejemplo.

Ministerio de Salud y Protección Social de Colombia. (2021). *Cáncer de cuello uterino*. Minsalud.Gov.Co. https://www.minsalud.gov.co/salud/publica/ssr/Paginas/Cancer-de-cuello-uterino.aspx

Mohammed A, Abdurahman F, & Abebe Y. (2021, June 29). *Single-Cell Conventional Pap Smear Image Classification Using Pre-Trained Deep Neural Network Architectures*. BMC Ingeniería Biomédica. https://bmcbiomedeng.biomedcentral.com/articles/10.1186/s42490-021-00056-6

Nanda, K., McCrory, D. C., Myers, E. R., Bastian, L. A., Hasselblad, V., Hickey, J. D., & Matchar, D. B. (2000). Accuracy of the Papanicolaou test in screening for and follow-up of cervical cytologic abnormalities: a systematic review. *Annals of Internal Medicine*, *132*(10), 810–819. https://doi.org/10.7326/0003-4819-132-10-200005160-00009

National Cancer Institute. (2021). *Pronóstico y tasas de supervivencia del cáncer de cuello uterino*. Cancer.Gov. https://www.cancer.gov/espanol/tipos/cuello-uterino/supervivencia

Pérez-González, O. A., & Aguilar-Lemarroy, A. (2015). *Patrón histológico del cáncer cérvico-uterino y su relación con el virus del papiloma humano*. Revista Médica Del Hospital General de México. https://revistamedica.com/patron-histologico-cancer-cervicouterino-virus-papiloma-humano/

Pérez, P., & Valente, M. (2018). *Fundamentos básicos del procesamiento de imágenes*. Facultad de Matematica, Astronomia, Fisica y Computación (UNC). https://www.famaf.unc.edu.ar/~pperez1/manuales/cim/cap2.html

Plissiti, M. E., Dimitrakopoulos, P., Sfikas, G., Nikou, C., Krikoni, O., & Charchanti, A. (2018). *SIPAKMED: A NEW DATASET FOR FEATURE AND IMAGE BASED CLASSIFICATION OF NORMAL AND PATHOLOGICAL CERVICAL CELLS IN PAP SMEAR IMAGES*. Ieee Xplore. https://ieeexplore.ieee.org/document/8451588/references#references

Pramanik, R., Biswas, M., Sen, S., de Souza Júnior, L., Papa, J., & Sarkar, R. (2022, June). *A fuzzy distance-based ensemble of deep models for cervical cancer detection*. Computer Methods and Programs in Biomedicine. https://www.sciencedirect.com/science/article/abs/pii/S0169260722001626

Rahaman, M., Li, C., Yao, Y., Kulwa, F., Wu, X., Li, X., & Wang, Q. (2021, July). *Deep Cervix: a deep learning-based framework for the classification of cervical cells using hybrid deep feature fusion techniques*. Computers in Biology and Medicine. https://www.researchgate.net/publication/353477753_DeepCervix_A_deep_learning-based_framework_for_the_classification_of_cervical_cells_using_hybrid_deep_feature_fusion_techniques

Rodrigo, J. A. (2017, April). *Máquinas de Vector Soporte (Support Vector Machines, SVMs)*. Cienciadedatos.Net. https://www.cienciadedatos.net/documentos/34_maquinas_de_vector_soporte_support_vector_machines#M%C3%A1quinas_de_Vector_Soporte

Schlagenhauf, T. (2022). *Linear Discriminant Analysis in Python | Machine Learning*. Javatpoint. https://python-course.eu/machine-learning/linear-discriminant-analysis-in-python.php

scikit-learn 1.2.2 documentation. (2021). *sklearn.manifold.TSNE*. Scikit-Learn 1.2.2 Documentation. Retrieved April 30, 2023, from https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html

SIPaKMeD. (2020). *Cervical Cancer largest dataset (SipakMed)*. Kaggle.

Wikipedia contributors. (2022a, November 6). *Koilocyte*. Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Koilocyte&oldid=1120416157

Wikipedia contributors. (2022b, November 23). *Vaginal cytology*. Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Vaginal_cytology&oldid=1093098546

Wikipedia contributors. (2022c, November 23). *Vaginal cytology*. Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Vaginal_cytology&oldid=1093098546