UNIVERSIDAD DE LOS LAGOS

DETECCIÓN DE PRESENCIA DE PARÁSITOS EN EXAMEN PARASITOLÓGICO SERIADO DE DEPOSICIONES CON VISIÓN POR COMPUTADORA

DEPARTAMENTO DE CIENCIAS DE LA INGENIERÍA

INGENIERÍA CIVIL EN INFORMÁTICA

CAMPUS PUERTO MONTT, CHILE

Diego Ignacio Muñoz Viveros

diegoignacio.munoz@alumnos.ulagos.cl

Profesor Guía: Joel Sebastian Torres Carrasco

Co-guía: Carlos Dupré Alvarado

14 de agosto de 2021

Agradecimientos

Gracias

Resumen

El examen parasitológico seriado de deposiciones es un examen realizado para la detección de presencia parasitológica en pacientes. El examen se lleva a cabo con la toma de muestras de deposiciones del paciente y posteriormente se hace observación de las mismas por medio de microscopio, en la realización de la observación se documenta la confirmación y clasificación de presencia parasitológica en orden de dar con un tratamiento certero, eficaz y acorde a las detecciones. Que el examen sea seriado refiere a las condiciones en las que estas muestras, tres muestras enfrascadas por separado, serán tomadas con la finalidad de tener muestras en distintos ciclos larvales.

El uso de la visión por computadora espera demostrar el aumento en la precisión y velocidad de la detección y clasificación de parásitos con objetivo de reducir incertidumbre y error humano introducido en la observación manual ejercida en el proceso de observación del examen parasitológico seriado de deposiciones. La visión por computadora es la combinación de tecnologías que permite a las computadoras generar inferencia respecto a imágenes estáticas o en movimiento emulando la visión humana por medio de un aprendizaje basado en conjuntos de datos utilizados como ejemplos iniciales de los cuales el modelo generado ajustará sus parámetros.

En la actualidad y en vista de los avances ejercidos en el área se reconoce como principal desafió la generación de un conjunto de datos de entrenamiento de los cuales se conozcan su grado de sesgo en la información que contenga de manera de evitar ajustes de un modelo muy débil o extremadamente limitado en sus capacidades de detección y clasificación cayendo en lo que es conocido como *Underfiting* y *Overfiting*.

Palabras Clave— Parasitólogia, Serializado, Deposiciones, Visión por Computadora, Modelo, Conjunto de Datos, *Underfiting*, *Overfiting*

Índice general

Capítulo 1

Introducción

En la actualidad en los laboratorios médicos existen gran cantidad de exámenes relacionados a la observación microscopica de muestras, estos exámanes son ralizados bajo observación humana de manera totalmente manual. Normalmente los exámenes que involucran esta observación microscopica suelen demandar bastante tiempo por muestra, esto multiplicado a la demanda habitual en el área de la salud nos da a conocer una problemática relacionada a la incapacidad de suplir la demanda en un tiempo pertinente.

Otra característica inerente a la observación está relacionada al error humano introducido en las observaciónes, resultando en observaciones con clasificaciones erroneas, esto concluyendo en el mal diagnostico médico. En ocaciones puede haber una gran diferencia entre las observaciones de distintos tecnólogos médicos sobre la misma muestra, esto en resumen, agrega gran incertidumbre a los análisis que normalmente solo puede ser resuelta con observaciones adicionales o complementarias.

La tecnología médica en el ámbito del conocimiento abarca el estudio e investigación que tiene como objetivo la aplicación de diferentes tipos de tecnologías para mejorar la salud de las personas durante su diagnóstico, desarrollo de la enfermedad y tratamiento aplicado. En el contexto médico, la tecnología médica es la rama de la salud que involucra a profesionales que se forman en los avances tecnológicos para aplicarlos a la medicina y las ciencias de la salud.

Para la consultoria de este trabajo se cuenta con la ayuda del laboratorio clínico CESFAM de Hornopiren, con la ayuda del tecnólogo médico y jefe del laboratorio clínico Carlos Dupré Alvarado.

El examen parasitológico seriado de deposiciones es un examen realizado para la detección

de presencia parasitológica en pacientes. El examen se lleva a cabo con la toma de muestras de deposiciones del paciente y, posteriormente, se hace observación de las muestras por medio de observación microscopica, en la realización de la observación se documenta la confirmación y clasificación de presencia parasitológica en orden de dar con un tratamiento certero, eficaz y acorde a las detecciones. La denominación seriado refiere a las condiciones en las que estas muestras, tres muestras enfrascadas por separado, serán tomadas con la finalidad de tener muestras en distintos ciclos larvales.

El problema es que dado lo extenuante de la observación y lo altamente susceptible a sesgo del observador el examen es muy sensible a perdida o mala clasificación de avistamientos llevando a malos diagnósticos médicos. En este trabajo se propone una solución inteligente para automatizar el examen parasitológico seriado de deposiciones para poder mejorar la calidad de los resultados y así mejorar la calidad de vida de las personas.

Para realizar este proyecto se pretende usar tecnologías que permitan aprovechar la experiencia de los tecnólogos médicos y utilizarla para crear mejores resultados. Es por ello que se pretende utilizar tecnologias de visión por computadora para automatizar la observación de muestras y a su vez reducir el grado de error mientras se apunta a una implementación rápida y simple.

La visión por computadora es la combinación de tecnologías que permite a las computadoras generar inferencia respecto a imágenes estáticas o en movimiento emulando la visión humana por medio de un aprendizaje basado en conjuntos de datos utilizados como ejemplos iniciales de los cuales el modelo generado ajustará sus parámetros. El uso de la visión por computadora espera demostrar el aumento en la precisión y velocidad de la detección y clasificación de parásitos con objetivo de reducir incertidumbre y error humano introducido en la observación manual ejercida en el proceso de observación del examen parasitológico seriado de deposiciones.

El desafío técnico encontrado está relacionado a la estructuración de un conjunto de datos de imágenes microscopicas que contengas la información suficiente para generar un modelo robusto y de altas prestación con gran capacidad de generalización para el examen parasitologico seriado de deposiciones.

Capítulo 2

Marco Teórico

2.1. Parasitología

2.1.1. Definición

La parasitología es la rama de las ciencias biologicas dedicada a el estudio de organismos, denominados parásitos, que dependen de otro para poder sobrevivir y que ocasionan grandes daños a las especies de las cuales dependen, relación llamada parasitismo.

La parasitología es una disciplina con aplicación en campos variados como medicina, farmacologia y veterinaria. Es utilizada en la investigación de parásitos que pueden producir enfermedades en plantas y animales con objeto de analizar, diagnosticar y posteriormente establecer un tratamiento óptimo para poder curarlas y erradicarlas.

Gran parte de los parásitos más dificil de tratar son los que se alojan en el interior del organismo, lo que puede ingresar por via oral o fluidos y gran parte de estos pueden alojarse en el sistema digestivo, principalmente en estomago e intestino.

En el contexto de la medicina, la área de **Tecnología Médica** en su especialización de parasitología esta encargada de la realización y analisis de exámenes con la finalidad de diagnosticar amenazas relacionadas a la disciplina.

2.1.2. Exámenes parasitologicos

Existen muchos tipos de análisis de laboratorio para diagnosticar enfermedades parasitarias. El tipo de análisis que solicite el médico se basará en sus signos y síntomas presentados durante la consulta médica, cualquier otra afección médica que pueda tener y sus antecedentes de viajes.

El análisis de laboratorio se lleva a cabo con las observaciones de muestras entregadas al laboratorio por el médico tratante. Estas muestras dependen de la busqueda de los parásitos sospechados y sus posibles ubicaciones, siendo estas muestras de la forma de sangre, heces, muestras urogenitales, esputo, aspirados o biopsias. La especificidad de los exámanes puede variar en la capacidad de detectar diferentes especies o realizar busquedas de manera particular.

Estos exámenes se pueden dividir en dos categorias:

- Invasivos: la adquisición de la muestra requiere intervención quirurgica algún tipo como las biopsias.
- No invasivos: la toma de la muestra presenta un método de obtención que no involucra una intervención invasiva al paciente como serían muestras de sangre o heces.

2.1.3. Procedimiento de exámenes

Para la realización de los exámenes se procede de las siguientes formas:

1. **Exámenes de muestra de sangre**: la muestra es tintada y analizada por goteo grueso y/o fino con un microscopio. El goteo fino es una forma de repartir la muestra en un portaobjeto [1] a manera de dejar una capa delgada y uniforme en la cual realizar observasiones, el goteo grueso por otro lado, consiste en soltar una gota de muestra de forma que la tensión superficial de la muestra mantenga su forma circular para dejar decantar las celulas contenidas en la muestra al fondo. La tinción es el proceso en el cual se suman compuestos a la muestra que reaccionan a componentes conocidos con el fin de teñir componente para facilitar la visualización.
2. **Endoscopia/Colonoscopia**: Conciste en la insersión en la boca (endoscopia) o el recto (colonoscopia) de una sonda con la cual el médico, normalmente un gastroenterólogo, examina

[1] Placa de acrílico trasparente usada para manejo de muestras para microscopio

las cavidades en busca de presencia parasitaria y/o lesiones de caracter aludible a presencia parasitaria.

3. **Exámenes parasitológico seriado de deposiciones**: consiste en el análisis de tres muestras seriadas [2] de heces tintadas con observación por microscopio . La observación se realiza por goteo fino.

4. **Resonancia Magnética (RM), Tomografía axial computarizada (TAC)**: Pruebas ralizadas para buscar enfermedades parasitarias que pueden provocar lesiones en los órganos.

2.1.4. Examen Parasitológico Seriado de Deposiciones

Como se menciona anteriormente, el examen parasitológico seriado de deposiciones es un examen que busca estudiar la presencia parasitaria intestinal por medio de observaciones de muestras de heces. La realización de este examen es llevado a cabo por un tecnólogo médico y su principal objetivo es documentar las observaciones realizadas.

La realización del examen en detalle se muestra a continuación:

1. Una vez se reconoce la necesidad de la realización del examen, el médico trantante pedirá la preparación de tres frascos de muestras con 10-20 ml de liquido fijador[3].

2. Se hace entrega al paciente de los frascos junto a una nota con las instrucciones para la toma de muestras, conservación de las mismas y precauciones a tomar.

3. Cuando se reciben las muestras y se valida su tamaño, los tiempos en los que se tomaron y la forma en que se conservaron, estas pasaran a ser entragadas al laboratorio donde seran almacenadas y/o examinadas.

4. Las muestras son retiradas del área de refrigeración, en el caso que se ecuentren en concervación, y se proceden a intruducir una porción de cada frasco en tres probetas diluidas en el liquido fijador.

[2] Muestras tomadas con intervalos de tiempo equidistantes con objetivo de muestrear sin que se pierdan ciclos larvarios evitando excluir avistamientos

[3] Liquido conservantes, normalmente formaldehído

5. Se centrifugan las muestras en orden de separar el material fecal del liquido fijador para poder extraerlo.

6. En caso de que el tecnólogo lo considere necesario, tintar la muestras con soluciones reactivas que resalten componentes especificos para facilitar las observaciones.

7. Dejar caer una gota de cada una de las muestras en un portaobjeto y esparcir para observación de goteo fino.

8. Hacer observación de las muestras con gran detalle en el microscopio y documentar hallazgos en caso de haber.

9. En caso de no detectar presencia parasitaria, documentar que no se observa presencia en la muestra.

Para la correcta realización del examen existen ciertas consideraciones expuestas a continuación:

- La toma de muestras debe avarcar un período mínimo de cinco días.
- La duración de la observación microscópica puede durar al rededor de 45 minutos y es dependiente de la habilidad y experiencia del tecnólogo.
- En caso de encontrar alguna anormalidad en la recepción de las muestras por parte del paciente, se debe pedir que este haga la toma nuevamente en frascos diferentes.
- La muestra que no se utilice durante la observación deber ser concervada en refrigeración.
- En caso de existir dudas sobre la observación de las muestras, se puede realizar una segunda observación.

2.2. Visión por Computadora

2.2.1. Definición

La visión por computadora es el campo de la inteligencia artificial (IA) que permite a los computadores y sistemas derivados a extraer información util de imagenes digitales, videos y

otras entradas visuales y tomar actiones o hacer recomendaciones basadas en esta información. Si la IA permite a los computadores pensar, la visión por computadora les permite ver, observar y entender [IBM, 2020].

La visión por computadora, así como muchas otras ramas de la disciplica de IA, logra la tarea de inferir gracias al aprendizaje de máquinas. El aprendizaje de maquinas consiste en el entrenamiento de modelos de IA en base a conjuntos de datos que harán que los distintos modelos aprendan a reconocer patrones y generen la lógica interna necesaria para permitirles inferir de manera correcta.

Para definir formalmente el problema de la visión por computadora [Prince, 2012] partimos tomando un dato visual x y lo usamos para inferir un estado del mundo w. El estado del mundo w puede ser de naturaleza continua (e.g. la posición tridimencional del modelo de un cuerpo) o discreto (e.g. la presencia o ausencia de un objeto particular). Cuando el estado es continuo, la inferencia es llamada regresión. Cuando el estado es discreto, la inferencia es llamada clasificación.

Esta definición nos permite definir el proceso de inferencia como $Pr(w|x)$, es decir, la probabilidad de que sierto estado del mundo w siendo que se dió x dato visual. Se puede notar rapidamente que a partir de esta definición solo se puede inferir sobre estados w conocidos. Esto en práctica se resuelve utilizando un estimador $\hat{w}$ que resulte suficientemente cercano.

Para resolver un problema de visión de esta naturaleza se necesitan tres componentes:

- Un modelo matemático que represente los datos visuales x y los estados del mundo w. El modelo representa una familia de posibles relaciones entre x, w y la relación particular es determinada por el modelo de parámetros θ.
- Un algoritmo de aprendizaje que nos permita ajustar los parámetros θ usando ejemplos de pares de entrenamiento $\{x_i, w_i\}$, donde sabemos ambas medidas y el estado relacionado.
- Un algoritmo de inferencia que tome una nueva observación x y use el modelo para calcular $Pr(w|x, \theta)$ sobre el estado del mundo w.

Los modelos que relacionan los datos x a el mundo w caen en una de dos categorias:

1. modelar la contingencia del estado del mundo en los datos $Pr(w|x)$ o

2. modelas la contingencia de los datos sobre el estado del mundo $Pr(x|w)$.

El primer tipo de modelos es denominado discriminativo. El segundo es denominado generativo; aquí, construimos un modelo de probabilidad sobre los datos y esto puede ser usado para generar nuevas observaciones.

Para efectos de este informe se estudiara el primer tipo de modelo.

2.2.2. Modelos de Clasificación

Al modelar $Pr(w|x)$, elegimos uns forma de distribución apropiada para $Pr(w)$ sobre los estados del mundo w y hacer la distribución de los parámetros una función de los datos x. Si el mundo de estados es continuo, lo definimos como una distribución normal con media μ en función de x [Fuchs, 2017].

El valor que nos entrega la función tambien en un conjunto de parámetros definidos, θ. Dado que la distribución dependen tanto de los datos x y los parámetros θ, escribimos que la función es $Pr(w|x,\theta)$ y nos referimos a ella como distribución posterior.

El objetivo del algoritmo de aprendizaje es el de ajustar los parámetros θ usando pares de entrenamiento $\{x_i, w_i\}$. Esto puede hacerse utilizando los enfoques de máxima verosimilitud, máximo a posteriori, o enfoques bayesianos .

2.2.3. Ajuste del Modelo Probabilístico

Nos referimos como aprendizaje al proceso de ajustar el modelo probabilístico al conjunto de datos $\{x_i\}$ ajustando los parámetros θ y es de vital importancia que el modelo pueda calcular la probabilidad de un nuevo dato $x*$ [Tiño, 2016].

Regresión Logística

Para los modelos de clasificación consideraremos la regresión logística [Geert, 2020], la cual independientemente de su nombre es un modelo que puede ser aplicado a clasificación. La regresión logística es un modelo discriminativo; nosotros seleccionamos una distribución de probabilidad sobre un estado del mundo $w \in \{0,1\}$ y hacemos estos parámetros concistentes en los datos observados x. Como los estados del mundo son binarios, lo describimos con una distribución de

Bernoulli y hacemos el parámetro λ (que indica la probabilidad de que el estado del mundo tome el valor $w = 1$) una medida en función de x.

Es importante destacar, no podemos simplemente hacer el parámetro λ una función linear $\phi_0 + \phi^T x$ de las medidas; una función linear puede retornar cualquier valor, pero el parámetro λ debe caer entre 0 y 1. Consecuentemente, primero debemos calcular la función linear y luego pasarla a través de la función sigmoidea que mapea el rango $[-\infty, \infty]$ a $[0, 1]$. El modelo final queda

$$Pr(\omega|\phi_0, \phi, x) = Bern_\omega[sig[\alpha]],$$

donde α es denominado la activación y es dado por la función linear

$$\alpha = \phi_0 + \phi^T x$$

y donde la función sigmoidea [Wood, 2020] es de la forma

$$sig(x) = \frac{1}{1 + e^{-x}}$$

Así como la activación α tiene a infinito esta función tiende a uno. Y así como tiende a menos infinito tiende a cero. Cuando α es cero, la función logística sigmoidea retorna un valor de un medio. Para datos unidimencionales de x, el conjunto de efectos de esta transformación es la de describir una curva sigmoidea relacionando x con λ. La posición horizontal de la sigmoidea es determinada por el lugar donde la función linear α cruza cero y su pendiente depende del gradiente ϕ_1.

Método Bayesiano

En el método bayesiano [Rennie, 2003, Backlund, 2017] dejamos de intentar estimar un solo valor de los parámetros θ y aceptamos lo obvio; abrán muchos valores de parámetros que serán compatibles con los datos. Calculamos la distribución de probabilidad $Pr(\theta|x_{1...I})$ sobre los parámetros θ basados en los datos $\{x_i\}_{i=1}^I$ usando el teorema de Bayes

$$Pr(\theta|x_{1...I}) = \frac{\Pi_{i=1}^I Pr(x_i|\theta) Pr(\theta)}{Pr(x_{1...I})}$$

Evaluando la distribución predictiva es más dificil para el caso bayesiano dado que no tenemos ningún modelo estimado pero en su lugar encontramos una distribución de probabilidad sobre posibles modelos. Así, calculamos

$$Pr(x*|x_{1...I}) = \int Pr(x*|\theta)Pr(\theta|x_{1...I})d\theta,$$

esto puede ser interpretado como: el termino $Pr(x*|\theta)$ es la predicción para un valor dado de θ. Así, la integral puede entenderse como la suma ponderada de los diferentes parámetros θ, donde la ponderación está dada por la distribución de probabilidad posterior $Pr(\theta|x_{1...I})$ sobre los parámetros (asumiendo que los parámetros son diferentes).

Los calculos de densidad predictiva para el bayesiano, máximo a posteriori u máxima verosimilitud puede unificarse si consideramos que estos dos últimos son distribuciones de probabilidades especiales sobre los parámentros donde toda la densidad está en $\hat{\theta}$. Más formalmente, los consideramos funciones delta centradas en $\hat{\theta}$. Una función delta $\delta(z)$ es una función que integra a uno, y retorna cero en todo su dominio excepto $z = 0$.

Entonces escribrimos

$$Pr(x*|x_{1...I}) = \int Pr(x*|\theta)\delta[\theta - \hat{\theta}]d\theta,$$

$$= Pr(x*|\theta),$$

finalmente vemos calculamos las probabilidades evaluando la probabilidad de los datos bajo el modelo con el que estimamos los parámetros.

Árboles de Decisión

Para implementar ese modelo [Rokach, 2007] comenzamos particionando es espacio de datos en distintas regiones y aplicamos diferentes clasificaciones a cada región.

El modelo de regresión ramificada logística tiene activaciones,

$$\alpha_i = (1 - g[x_i, \omega])\phi_0^T x_i + g[x_i, \omega]\phi_1^T x_i$$

El término $g[x, \omega]$ es una función de activación que retorna un número entre 0 y 1. Si esta

función de activación retorna 0, entonces la activación será $\phi_0 x_i$, por otro lado si retorna 1, la activación será $\phi_1 x_i$. Si la activación retorna un valor intermedio, entonces la activación será la suma ponderada de los dos componentes. La propia función de activacion depende de los datos x_i y toma parámetros ω. Este modelo induce un complejo límite de decisión no linear donde las dos funciónes lineares $\phi_0 x_i$ y $\phi_1 x_i$ son especializadas en diferentes regiones del espacio de datos.

La función de activación puede tomar muchas formas, pero una obvia posibilidad es la de usar un segundo modelo de regresión linear. Vale decir, calculamos una función linear $\omega^T x_i$ de los datos que son pasados a través de una sigmoidea logística, es decir,

$$g[x_i, \omega] = sig[\omega^T x_i]$$

Para que este modelo aprenda debemos maximizar $L = \Sigma_i \log[Pr(w_i|x_i)]$ una probabilidad logaritmica de los pares de datos de entrenamiento $\{x_i, w_i\}_{i=1}^I$ con respecto a todos los parámetros $\theta = \{\phi_0, \phi_1, w\}$. Normalmente esto puede lograr usando métodos de optimización no linear.

Podemos extender esta idea para crear una estructura de árbol jerárquica anidando funciones de activación. Por ejemplo,

$$\begin{aligned} \alpha_i = (1 - g[x_i, w]) \left[\phi_0^T x_i + (1 - g[x_i, w_0])\phi_{00}^T x_i + g[x_i, w_0]\phi_{01}^T x_i\right] \\ + g[x_i, w] \left[\phi_1^T + (1 - g[x_i, w_1])\phi_{01}^T x_i + g[x_i, w_1]\phi_{11}^T x_i\right] \end{aligned} \tag{2.1}$$

Esto es un ejemplo de árbol de clasificación.

Para aprender los parámetros $\theta = \{\phi_0, \phi_1, \phi_{00}, \phi_{01}, \phi_{10}, \phi_{11}, w, w_0, w_1\}$ podemos tomar una aproximación incremental. En la primera etapa ajustamos la parte superior del árbol, ajustando los parámetros w, ϕ_0, ϕ_1. Luego ajustamos la rama de la izquierda, ajustando los parámetros $w_0, \phi_{00}, \phi_{01}$ y así con la rama de la derecha, ajustando $w_1, \phi_{01}, \phi_{11}$ y así sucesivamente.

Bosque de Árboles Aleatorio

Esta idea se volvió popular en la problemas de clasificación multiclase usando la idea de árbol de decisión definida anteriormente.

Un bosque aleatorio [Ho, 2016] es una colección de árboles aleatorios, cada uno de estos usa

un conjunto de funciónes elegidas de manera aleatoria. Cuando se promedian todas las probabilidades de cada árbol, es decir, $Pr(w*|w*)$ predicha por cada árbol, se produce una clasificación mucho más robusta. Una forma de pensarlo es aproximarlo al método bayesiano; donde construimos una respuesta final tomando una suma ponderada de las predicciones propuestas por los distintos conjuntos de parámetros.

2.2.4. Conjunto de Datos

El manejo de los conjuntos de datos [Byjus, 2012b] es una de las tareas más importantes a la hora de diseñar un modelo de IA, ya que la información y como esté clasificada marcará como el modelo podrá desempeñarse en un entorno de uso general o específico que no contenga información vista con anterioridad.

La estructura general de un conjunto de datos en modelos de visión por computadora suele consistir en un conjunto de imagenes o videos junto a etiquetas que muestran la información contenida en estas, normalmente en la forma de coordenadas que delimitan el área en donde se localiza en la imagen el objeto a identificar y/o clasificar, junto a la etiqueta que identifíca que es o que acción esta ocurriendo.

Es importante contar con un conjunto de datos que contenga gran cantidad de ejemplos relacionados a la tarea que se espera lograr, esto con objetivo de generalizar los patrones e información obtenida por parte del modelo durante el proceso de entrenamiento. La cantidad de información que se requiera estará dada por la complejidad en el modelo a entrenar y la complejidad de la tarea a desarrollar, y el análisis de estos parámetros dependerá del equipo de desarrollo. Cuando en entrenamiento carece de un conjunto de datos contundente o el modelo es muy simple para la tarea ocurre *underfiting*, esto es, cuando el modelo realiza inferencias escazas y/o erroneas.

Para un hacer buen provecho de la información existen diferentes tratamientos realizables a los conjuntos de datos como lo son el limpieza de datos, aumento de datos y el balanceo de datos.

Limpieza de Datos

La limpieza de datos es un procedimiento que nos ayuda a filtrar datos erroneos y manipular información faltante para no causar una mayor perdida de datos [Byjus, 2012a]. Este procedimiento es descrito como la intuición de los datos, dado que no existen muchas más formas de entender

un conjunto de datos que analisandolo de manera manual. Generalmente el manejo de la limpieza de datos suele ser una de las partes más demandantes en el proceso de desarrollo de un modelo de inteligencia artificial.

Los problemas que se presentan en conjuntos de datos que debes ser solbentados son:

- **Datos faltantes**: Existen situaciones en las que no todos los datos de una entrada tendran asignados un valor. Si bien es factible eliminar todas las entradas que contengan datos perdidos, en la mayoria de los casos esto termina descartando excesiva cantidades de entradas (hay conjuntos de datos que pueden tener datos perdidos en todas sus entradas).

  Una alternativa para cuando los datos faltantes tienen estar agrupados en su mayoria en un campo en particular es eliminar el campo por completo. Esto valido dado que el campo de por si mostraba no tener mucha utilidad.

  La forma más común de manejar datos faltantes es usar un valor por defecto que en ultima instancia servirá para que el modelo reconozca que existen situaciones donde los datos no se encuentran disponibles. Esto es basicamente porque siempre se puede extraer información del echo de que no existen entradas en un campo en particular [Tatman, 2020a].

- **Distribución de datos inconsistentes**: Esto ocurre cuando la escala o el formato de los datos no esta bien definido en los campos, vale decir, tener números en campos que describen texto o números en rangos que no corresponden.

  Cuando se habla de inconsistencia en tipo de datos solo basta con reconocer el tipo de dato del campo para hacer la corrección del resto de los datos, revisando que la transformación de los datos sea consistente al objetivo final.

  Cuando la inconsistencia corresponde a rangos se implementan dos alternativas, escalado y normalización.

  El escalado corresponde a una transformación en el rango de los datos, esto combierte los datos tomando el mínimo y máximo valor en los campos y escalandolos a valores entre 0 y 1.

  La normalización es una transformación que convierte la distribución de los datos en una distrubución normal Gaussiana. Este método es utilizado en datos destinados a modelos de predicción estadística [Tatman, 2020c].

- **Datos inconsistentes**: Es el tipo de dato que requiere manipulación acorde a las necesidades, estos datos pueden ser desde texto que registra entradas diferentes por tener su versión iniciando con mayusculas y su versión en minusculas hasta tipos de datos diferentes en las mismas columnas.

  Para la reparación de estos datos se debe hacer estudio de los datos y la información que se espera de ellos y es totalmente sujeta a quien manipula los datos [Tatman, 2020b].

Aumento de Datos

Muchos de los desarrollos de modelos actuales presentan una característica en común, fueron entrenados con cantidades de datos que pasan el orden de magnitud de los cientos milos o millones. Esto es un importante dato a considerar cuando se tienen modelos que no cuentan con un conjunto de datos lo suficientemente amplio.

Para esto es que se hacen estudios respecto el aumento artificial de datos, que consisten en resumidas cuentas en el uso de datos pre-existentes en el conjunto de datos actual para creas nuevas entradas que representen información similar. Hay estudios que muestran aumentos de rendimiento sobre el 30 % [K, 2020] sobre modelos que no la utiliza.

Aumento de datos numéricos

Las técnicas de aumento utilizadas [Khare, 2011] en las aplicaciones de aprendizaje profundo dependen del tipo de datos. Para aumentar los datos numéricos simples, son populares técnicas como SMOTE o SMOTE NC. Estas técnicas se utilizan generalmente para abordar el problema del desequilibrio de clases en las tareas de clasificación.

Para los datos no estructurados, como las imágenes y el texto, las técnicas de aumento varían desde simples transformaciones hasta datos generados por redes neuronales, en función de la complejidad de la aplicación [Nayak, 2020].

Balanceo de Datos

En cierto tipo de conjuntos de datos ocurre que cierta concentración de datos sobrepasa de sobre manera a otra, causando desbalanceo de datos. Una forma de sobre llevar esto es particio-

nando el segmento de datos más abundante para poder representar toda la información con el mismo nivel de importancia.

Es importante reconocer las desventajas a lo que esto puede con llevar; esto hace dejar de lado gran cantidad de datos, esto es proporcional al conjunto que represente la minoria de datos en el conjunto, esto culminando en *underfiting*.

2.2.5. Procesamiento de Imágenes y Extracción de Características

Es importante saber reducir la complejidad del análisis de imagenes sabiendo reconocer que tipo de solución necesitamos para nuetro problema. Para ello existen multiples transformaciónes que permiten simplificar la entrada visual que recibirá el modelo en orden de facilitar la tarea de extracción de características claves [Prince, 2012].

Transformación por Pixel

Es la transformación más directa de imagenes [Wang, 2020], consisten en la modificación de pixeles individuales asumiendo la estructura bidimensional. Se dice que la transformación modifica el pixel p_{ij} donde i, j son las coordenadas del pixel en el arreglo de la imagen P.

Escala de Grises

Es convertir la imagen de entrada a color a su versión a escala de grises. Este filtro es especialmente util para reducir el arreglo de pixeles de valores rgb a un simple valor que representa el nivel de blanco del pixel, esto es útil en los casos donde los colores no son relevantes en la extracción de caracteristicas [Wang, 2020].

Filtros Lineales

Son transformaciónes de pixeles x_{ij} que consideran una suma ponderada del valor de los pixeles abyacentes [Ludwig, 2020], es decir, siendo una imagen P a la cual se aplicará un filtro denominado *kernel* F, el cual tendrá entradas f_{mn} donde $m \in \{-M, M\}$ y $n \in \{-N, N\}$. Finalmente

quedamos con un filtro de la forma

$$x_{ij} = \Sigma_{m=-M}^{M} \Sigma_{n=-N}^{N} p_{i-m,j-n} f_{m,n}$$

2.3. Estado del Arte

2.4. Objetivo General

Crear un modelo de aprendizaje de maquinas para la detección de y clasificación de parásitos en el análisis parasitológico de seriado de deposiciones.

2.5. Objetivos Específicos

1. Recopilar y estructurar un conjunto de datos de imágenes de muestras para el entrenamiento y testeo del modelo de detección de parásitos.
2. Definir un modelo de detección de parásitos para la automatización de resultados del examen parasitológico seriado de deposiciones basado en visión por computadora.
3. Analizar y validar la calidad de las predicciones de la detección de parásitos en las muestras, a través de experimentación con el conjuntos de datos recopilado y con muestras obtenidas desde procedimientos reales del examen parasitológico seriado de deposiciones, utilizando imagenes microscopicas.

2.6. Metodología

1. Recopilar y estructurar un conjunto de datos de imágenes de muestras para el entrenamiento y testeo del modelo de detección de parásitos.
    a) Revisión de conjunto de datos disponibles relacionados al examen parasitológico seriado de deposiciones.

b) Definir una estructura estandar para almacenar imagenes microscopicas de examen parasitológico seriado deposiciones.

c) Implementar un procedimiento de extracción y limpieza de imagenes de examen parasitológico seriado de deposiciones.

d) Realizar un procedimiento de análisis exploratorio de datos.

2. Definir un modelo de detección de parásitos para la automatización de resultados del examen parasitológico seriado de deposiciones basado en visión por computadora.

   a) Adoptar una metodología (elección de experimento, metricas, configuración y *kfolds*) de proyectos de visión por computadora para relizar una comparación de la calidad de los modelos.

   b) Proponer distintos modelos para la detección de parásitos en imagenes microscopicas.

   c) Seleccionar el o los modelos con mejor calidad de resultado.

3. Analizar y validar la calidad de las predicciones de la detección de parásitos en las muestras, a través de experimentación con el conjuntos de datos recopilado y con muestras obtenidas desde procedimientos reales del examen parasitológico seriado de deposiciones, utilizando imagenes microscopicas.

   a) Implementar la metodología elegida o adaptada para el tratamiento de imagenes microscopicas.

   b) Implementar el procedimiento de predicción de parásitos detectada en las imagen microscopicas.

   c) Implementar el reporte final de parásitos detectados en las muestras a través del modelo de visión por computadora.

   d) Recopilación de nuevas muestras, sus resultados, su validación y la retroalimentación del tecnologo médico.

2.6.1. Planificación

Figura 2.1: Carta Gantt para desarrollo del proyecto

Bibliografía

[Backlund, 2017] Backlund, G. (2017). *An Overview of the Data Augmentation Algorithm*. Department of Statistics, University of Florida.

[Byjus, 2012a] Byjus (2012a). Data sets meaning. Datasets.

[Byjus, 2012b] Byjus (2012b). What is data management? Data Managment.

[Fuchs, 2017] Fuchs, K. (2017). Machine learning: Classification models. Classification Models.

[Geert, 2020] Geert, R. (2020). Logistic regression. Logistic Regression.

[Ho, 2016] Ho, T. K. (2016). *Random Decision Forests*. AT&T Bell Laboratories.

[IBM, 2020] IBM (2020). What is computer vision? Computer definition by IBM.

[K, 2020] K, A. (2020). Understanding data augmentation | what is data augmentation & how it works? Data Augmentation.

[Khare, 2011] Khare, K. (2011). A spectral analytic comparison of trace-class data augmentation algorithms and their sandwich variants. Spectral Analytic and Comparations.

[Ludwig, 2020] Ludwig, J. (2020). *Image Convolution*. Portland State University.

[Nayak, 2020] Nayak, R. (2020). Data augmentation in natural language processing for text classification. Improve the performance of your model by generating data on the go.

[Prince, 2012] Prince, S. (2012). *Computer vision: models, learning and inference*. Cambridge University Press.

[Rennie, 2003] Rennie, J. D. (2003). *Tackling the Poor Assumptions of Naive Bayes Text Classifiers*. Artificial Intelligence Laboratory; Massachusetts Institute of Technology; Cambridge, MA 02139.

[Rokach, 2007] Rokach, L. (2007). *Decision Trees*. Tel-Aviv University.

[Tatman, 2020a] Tatman, R. (2020a). Handling missing values. Missing Data.

[Tatman, 2020b] Tatman, R. (2020b). Inconsistent data entry. Inconsistent Data.

[Tatman, 2020c] Tatman, R. (2020c). Scaling and normalization. Scaling and Normalize Data.

[Tiño, 2016] Tiño, P. (2016). *Probabilistic Modelling in Machine Learning*. Birmingham University, UK.

[Wang, 2020] Wang, Y. (2020). *Geometric Transformations: Warping, Registration, Morphing*. Polytechnic University, Brooklyn.

[Wood, 2020] Wood, T. (2020). What is the sigmoid function? Sigmoid Function.