
UNIVERSIDAD DEL VALLE
ESCUELA DE INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

# SISTEMA VISUAL AUTOMÁTICO DE CLASIFICACIÓN DE SEMILLAS DE FORRAJES PARA BANCOS DE RECURSOS GENÉTICOS

*Research project proposal supervised by*

MARIA PATRICIA TRUJILLO URIBE, PHD

*and submitted by*

DIEGO FERNANDO GONZALEZ MONROY

*to fulfil the requirement of*

MAESTRÍA EN INGENIERÍA CON ÉNFASIS EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

*26th May 2018*

# Abstract

El banco de recursos genético del CIAT se encarga de conservar la biodiversidad de los cultivos de Frijol y Forrajes, los cuales son de gran interés a nivel mundial, por tanto, es de importancia conservar las semillas de mejor calidad que conserven las características morfológicas y fenotípicas correspondientes a la semilla original.

En este proyecto de grado se propone desarrollar un modelo de clasificación que se ajuste a las necesidades del banco de recursos genéticos del CIAT, para esto será necesario identificar cuáles son los modelos de clasificación existentes y sus aplicaciones, teniendo en cuenta que el banco de recursos genéticos cuenta con recursos limitados, no se tienen grandes conjuntos de datos en cuanto a imágenes y la clasificación actualmente es realizada por personal experto que ha adquirido una gran experiencia a través de los años.

***Keywords*—** Clasificación, Reconocimiento de patrones, Procesamiento de imágenes, Bancos de recursos genéticos, Fenotipo, Calidad, kernel.

# Contents

# List of Figures

# Chapter 1

# Introduction

El Centro Internacional de Agricultura Tropical (CIAT) es uno de los 15 centros de investigación del CGIAR. La misión del CIAT es buscar reducir el hambre y la pobreza del mundo por medio de una agricultura eco-eficiente, es decir una agricultura competitiva, rentable, sostenible y accesible.

El Programa de Recursos Genéticos (PRG) del CIAT es el área encargada de conservar, distribuir y garantizar la biodiversidad de las colecciones de Frijol, Forrajes y Yuca (37987, 23140 y 6643 respectivamente). Para llevar a cabo su función es necesario realizar una serie de procesos tales como: introducción, regeneración/multiplicación, caracterización, cosecha, recepción y almacenamiento. En este último es necesario verificar y clasificar la calidad de las semillas para garantizar la biodiversidad de las colecciones.

Para el PRG la verificación y clasificación de las semillas es un proceso de vital importancia puesto que es necesario identificar las semillas de mejor calidad, que cumplan con unas características específicas de cada variedad, teniendo en cuenta que el PRG conserva más de 60.000 variedades (entres Frijol y Forrajes) y por cada variedad se debe seleccionar un promedio de 4600 semillas.

El proceso de verificación y clasificación de semillas actualmente se realiza seleccionando un promedio de 6000 semillas por cada variedad, las cuales deben ser clasificadas y analizadas manualmente por personal experto, quienes deben identificar las características morfológicas y fenotípicas de cada material de tal manera que seleccionen las semillas de mejor calidad, algunas de las características que busca identificar el personal experto son: intensidad del color, manchas, mezclas o segregaciones, perforaciones, tamaños, malformaciones, textura, forma, grosor y peso. Durante la clasificación por parte del personal experto se verifica visualmente la semilla consultando por medio del código de la variedad en un ordenador la imagen o haciendo uso del catálogo de imágenes físico. Este proceso de clasificación de semillas puede tardar entre 15 y 20 min para 6500 semillas de una sola variedad, dependiendo de la limpieza previa, tamaño y calidad de la semilla cosechada.

Después de realizar el proceso de clasificación de las semillas durante la semana, el último día se procede a realizar una comprobación visual de todas las semillas clasificadas, la cual es realizada por la persona más experimentada en la clasificación de semillas en los cultivos de frijol y forrajes, esta persona debe verificar cerca de 50

o 60 variedades cada una con un promedio de 4600 semillas pre-seleccionadas el último día de la semana.

Por tanto, se realiza una investigación de los modelos de clasificación de granos que presentan una gran precisión, la cual en algunos casos puede llegar a más del 90%, pero para esto tienden a usar grandes cantidades de datos de los cuales los bancos de recursos genéticos no disponen, a diferencia de las industrias. En este proyecto de maestría se propone desarrollar un prototipo que tenga en cuenta aspectos como: adquisición de imágenes, modelos de clasificación, conocimiento de los expertos, características morfológicas y fenotípicas. De manera que este pueda ser entrenado con una cantidad reducida de imágenes (menor a 6500) y que brinde una precisión mayor al 80% *Plant Production and Protection Division: Las Normas para Bancos de Germoplasma de Recursos Fitogenéticos para la Alimentación y la Agricultura* (n.d.) .

## 1.1. Research Problem

El Programa de Recursos Genéticos del CIAT conserva las colecciones de Frijol y Forrajes. Para lograr conservar la biodiversidad de las colecciones es necesario realizar el proceso de verificación y clasificación de semillas con un alto grado de precisión y calidad, pero en ocasiones esto es muy dispendioso puesto que el personal requiere de una gran experiencia y habilidad para identificar y seleccionar las semillas de mejor calidad. Lo cual hace la clasificación un proceso muy subjetivo.

El proceso de verificación y calificación de semillas se lleva a cabo diariamente en el área de conservación y se realiza manualmente por el personal experto, que ha adquirido el conocimiento y ha aprendido a clasificar las semillas de acuerdo a las características de cada variedad. Durante el proceso de clasificación se tienen en cuenta muchas variables que permite identificar las semillas adecuadamente, entre estas tenemos: altura, ancho, grosor, color de oxidación, color original, manchas, textura, forma, perforaciones y humedad. En caso de dudas durante la clasificación se compara la semilla con las imágenes de la semilla original que se encuentran en un computador o en un catálogo, ocasionando que la clasificación tome más tiempo ya que se debe buscar manualmente el nombre de la variedad. Además, en ocasiones las semillas a clasificar tienen un tamaño pequeño (1 mm - 3 mm), por lo cual se hace uso de una lupa o un sistema de visión 2D que permiten identificar adecuadamente estas semillas.

Sin embargo la lupa obliga al personal a esforzarse en busca del enfoque adecuado ocasionando que se adquieran posturas inadecuadas durante tiempos prolongados afectando la visión y estado físico. El sistema de visión con video cámara (Ergo Visión System) que permite visualizar la imagen en 2D de las semillas, a este sistema se le adaptó una bomba de succión que permite extraer con mayor precisión

y agilidad las semillas. Este sistema de visión depende del brillo, el aumento de la cámara y la velocidad del riel por el cual pasan las semillas, los cuales son regulados por el usuario manualmente.

## 1.2. Research Question

Cómo seleccionar en forma automática las semillas más aptas para conservación con pocos datos (<6500 imágenes) que brinde una precisión mayor al 80% teniendo en cuenta características morfológicas y fenotípicas que representan a las semillas de mejor calidad.

## 1.3. Justification

Los bancos de recursos genéticos cuentan con una gran diversidad de semillas de distintos cultivos como lo son arroz, maíz, trigo, frijol, forrajes entre otros, la misión de estos bancos es conservar y garantizar la diversidad genética de cada una de estas variedades. Cada cultivo contiene miles de semillas como es el caso del Banco de Recursos Genéticos del CIAT, que conserva 37987 variedades de frijol, 23140 variedades de forrajes y 6643 variedades de yuca, cada una de estas variedades deben cumplir con unos estándares en cuanto a cantidad de semillas, calidad y viabilidad, los cuales son constantemente monitoreadas para una adecuada conservación.

La cantidad de semillas es de gran importancia para la conservación puesto que, para obtener la cantidad de semillas necesarias a conservar, los materiales deben salir a campo a regeneración o multiplicación y estas pueden tardar de 2 meses hasta años en completar la cantidad de semilla suficiente a conservar, por tanto, la cantidad de semillas es limitada y se debe lograr una cantidad suficiente de acuerdo a la variedad para garantizar su conservación. La calidad es otra propiedad importante que deben tener las semillas puesto que esta garantiza que estas contaran con las características morfológicas y fenotípicas de la semilla original, adicionalmente deben contar con las mejores cualidades de una semilla sana (color, forma, tono, textura).

Mantener estos estándares de calidad y cantidad tienden a ser dispendiosos, costosos y difíciles de llevar a cabo puesto que en el caso del cultivo de forrajes las semillas pueden llegar a medir milímetros, lo cual dificulta su clasificación, selección y conteo, por tanto, esto genera un cuello de botella en los tiempos de trabajo, afecta la salud de los trabajadores y se presenta subjetividad en este proceso.

Se realizó una revisión del estado del arte de los distintos modelos de clasificación para cultivos de granos y no se encontraron aplicaciones enfocadas a los cultivos de forrajes, tampoco se encontró información acerca de cuál sería la cantidad mínima de imágenes para entrenar estos modelos puesto que se basaban en grandes

cantidades de datos. la mayoría de las aplicaciones han sido desarrolladas para industrias de granos que trabajan con toneladas de semillas como el arroz, cebada, trigo, etc. Se han encontrado relativamente pocos modelos de clasificación Hansen et al. (2016) desarrollados para contextos de bancos de recursos genéticos.

## 1.4. Objectives

### 1.4.1. General objective

Desarrollar de un sistema de inspección visual para la clasificación de forrajes.

### 1.4.2. Specific objectives

1. Definir un protocolo de captura de imágenes de semillas de forrajes.
2. Determinar el conjunto de características visuales que representen la morfologia y fenotipos de buena calidad de la variedad seleccionada.
3. Construir un clasificador usando las características seleccionadas.
4. Desarrollar un prototipo de software del sistema de inspección visual.

## 1.5. Proposal approach

En este proyecto de maestría se propone desarrollar un modelo de clasificación que reúna el conocimiento adquirido por los expertos, el conocimiento extraído de las imágenes y el conocimiento de las especies, de manera que se permita llevar a cabo la automatización del proceso de clasificación para semillas de tamaños pequeños (1 mm - 3 mm) y con una cantidad reducida de imágenes menor a 6500.

# Chapter 2

# Referential framework

## 2.1. Conceptual Framework

### Programa de recursos genéticos

Es el banco de germoplasma del CIAT donde las plantas (en forma de semillas o plántulas en tubos de ensayo) son catalogadas, conservadas para el largo plazo y puestas a disposición para distribución. Los bancos de germoplasma no son simplemente repositorios de la biodiversidad; ellos ayudan a los mejoradores e investigadores en la selección de materiales apropiados para el mejoramiento genético y la investigación. Además, proporcionan directamente semillas o material de siembra para los agricultores *Conservación y uso de cultivos* (n.d.).

### Características morfológicas

La caracterización morfológica de recursos filogenéticos es la determinación de un conjunto de caracteres mediante el uso de descriptores definidos que permiten diferenciar taxonómicamente a las plantas. Algunos caracteres pueden ser altamente heredables, fácilmente observables y expresables en la misma forma en cualquier ambiente. Las características morfológicas se utilizan para estudiar la variabilidad genética, para identificar plantas y para conservar los recursos genéticos Hernández-Villareal (2013).

### Características fenotípica

Una característica fenotípica es cualquier característica o rasgo observable de un organismo, como su morfología, desarrollo, propiedades bioquímicas, fisiología y comportamiento. Los fenotipos resultan de la expresión de los genes de un organismo, así como de la influencia de los factores ambientales, y de las posibles interacciones entre ambos *Fenotipos - Fenotipo .com* (n.d.).

## Calidad fitosanitaria

Es garantizar la trasferencia de materiales o semillas libre de patógenos y plagas las cuales pueden verse reflejadas en las características sanitarias y fisiológicas de las semillas Cuervo et al. (2016).

## Machine learning

El término Machine learning hace referencia a la detección automatizada de patrones de un amplio conjunto de datos. En las últimas décadas se ha convertido en una herramienta muy importante en la mayoría de tareas que requieren la extracción de información de grandes conjuntos de datos. Constantemente se están desarrollando herramientas o software basado en máquinas de aprendizaje como lo son los motores de búsqueda, carros autónomos, software anti-spam, reconocimiento de rostros, entre otros.

Existen distintos tipos de aprendizaje supervisado, no supervisado, paradigmas de aprendizaje activo y pasivo, en línea, protocolos de aprendizaje por lote, entre otros Shalev-Shwartz and Ben-David (2014).

## Support Vector Machine (SVM)

Es uno de los más conocidos métodos de Machine learning para clasificación, regresión y otras tareas de aprendizaje, este método basado en un grupo de muestras clasificadas (o conjunto de datos de entrenamiento previo) tiene la capacidad de predecir la clasificación a la que pertenece una nueva muestra ubicándola en un punto en el hiperplano de acuerdo a los puntos del entrenamiento previo Wang (2005).

## k-nearest neighbor (KNN)

Algoritmo que memoriza un conjunto datos de entrenamiento y luego predecir la etiqueta de cualquier clase nueva basado en las etiquetas de sus k vecinos más cercanos en el conjunto de entrenamiento. La razón detrás de este método se basa en la suposición de que las características que se utilizan para describir los puntos de dominio son relevantes para su etiquetado de una manera que hace que los puntos cercanos probablemente tengan la misma etiqueta. Además, en algunas situaciones encontrar un vecino más cercano puede hacerse muy rápido en un grupo de datos muy grande Beyer et al. (1999).

### Linear Discriminant Analysis (LDA)

Es más comúnmente usado como técnica de reducción de dimensionalidad en el paso de pre- procesamiento para aplicaciones de clasificación de patrones y aprendizaje de máquinas. El objetivo es proyectar un conjunto de datos en un espacio de menor dimensión con una buena separabilidad de clase para evitar el sobre ajuste y también reducir los costos computacionales, aunque en ocasiones es muy efectivo en tecnicas de clasificasión entre dos clases Mika et al. (1999).

### Naive-Bayes classifier

El clasificador Naive Bayes es una demostración clásica de cómo las suposiciones generativas y las estimaciones de parámetros simplifican el proceso de aprendizaje de manera que se Asume que todas las características son condicionalmente independientes dadas la etiqueta de la clase. A pesar de que esto es generalmente falso (ya que las características suelen tener dependencias), el modelo resultante funciona correctamente Zhang (2004).

### Artificial Neural Network (ANN)

Las redes neuronales artificiales surgen a partir del conocimiento del funcionamiento biológico de las neuronas, esto permite desarrollar un modelo de aprendizaje basado en neuronas (perceptro) las cuales permiten aprender por medio de un conjunto de datos de entrenamiento y funciones de activación (ReLu) que posteriormente son optimizadas por medio de técnicas de optimización o backpropagation Basheer and Hajmeer (2000).

## 2.2. State of the Art

Existen gran variedad de modelos de clasificación aplicados a distintas áreas de reconocimiento de patrones, en esta revisión de literatura se han encontrado modelos de clasificación (ANN, LDA, KNN, Fuzzy logic, etc.) aplicados a diferentes tipos de granos con el fin de poder discriminar entre variedades, granos defectuosos, granos de mejor calidad, desechos, entre otros. Como se muestra en la tabla 1.

| Modelos de clasificación aplicados a semillas | | | | | | |
|---|---|---|---|---|---|---|
| Clasificador | Cultivo | Precisión | Imág | Año | Ref | Descripción |
| ANN | Frijol | >90.6% | 511 | 2007 | 2007 | Clasificación de frijol por color (Blanco, Amarillo verdoso y Negro) y daño presente. |
| ANN y LDA | Cebada | >86% y>65% | 13000 | 2015 | 2015 | Clasificación de 11 variedades de cebada basados en características de color, forma y textura. |
| ANN | Pimienta | >84.94% | 832 | 2016 | 2016 | Clasificación entre 8 variedades de acuerdo a su conjunto de características. Identificación de variedades de frijol de acuerdo a color. |
| ANN y SVM | Pistacho | >99% y >99% | 850 | 2017 | 2017 | Clasificación entre 5 clases de pistacho basado en color y forma. |
| SVM | Cereal | >90% | | 2009 | 2009 | Clasificación de 14 clases entre cereales, cornezuelo de centeno, granos no sanos y quemados y otras impurezas |
| SVM | Maiz | >95% | 14000 | 2011 | 2011 | Clasificación de 14 clases una buena y 13 defectuosas a partir de color y textura. |
| KNN y LDA | Trigo | >93% y >91% | 4000 | 2013 | 2013 | Clasificación de 4 clases una buena y 3 tipos de daños basado en características morfológicas, textura y EFD. |
| KNN | Arroz | >93% | 600 | 2015 | 2016 | Clasificación de 20 variedades de arroz. |
| CNN | Frijol | >95% | 866 | 2016 | 2016 | Clasificación de venas foliares de las hojas de tres variedades de leguminosas. |
| Otros | - | - | - | 2006-2015 | 2006 | Revisión aplicada a alimentos (granos de cultivos, frutas, plantas y carnes). |

La anterior tabla muestra algunos de los modelos de clasificación más comunes hoy en día, los tipos de granos y la precisión con la que se han llevado a cabo las respectivas investigaciones.

La gran mayoría de los procesos de automatización para el reconocimiento de patrones en imágenes de granos con llevan a la automatización y estandarización de los métodos de adquisición de las imágenes, puesto que estos generan un gran impacto en la precisión de los modelos de clasificación. Algunos aspectos que se deben tener en cuenta para mejorar la calidad de las imágenes y tener mayor precisión en estos métodos son: tipo de imagen (RGB o Multi-espectral), distancia estandarizada del dispositivo fotográfico, pixeles, brillo, sombras, imágenes sin colisiones, etc.

Adicionalmente se encontró que en Jayas and Singh (2012) se revisan los dispositivos de captura de imágenes (video o foto) que permiten obtener una mayor cantidad de información donde se mencionan aplicaciones de Rayos X, Color (RGB), Hiperespectral o Multiespectral y Térmicas, enfocadas a la clasificación de calidad de los granos. En este artículo se resalta las imágenes hiperespectrales debido a la cantidad de información que se puede obtener con respecto a las otras técnicas al igual que se recomienda su uso en aplicaciones de selección de calidad. También se encontró aplicaciones de imágenes multiespectrales para clasificación de variedades de semillas de arroz en el banco de recursos genéticos IRRI Hansen et al. (2016) y variedades de semillas de tomate Shrestha et al. (2015), con una precisión mayor del 93% y 82% respectivamente.

## Tecnicas de clasificación

### 2.2.1. Artificial Neural Networks(ANN)

Las ANN Jain et al. (1996) han sido uno de los enfoques que se encuentran en auge durante los últimos años, inicialmente fueron mencionadas por McCulloch et al McCulloch and Pitts (1943), quienes desarrollaron un modelo matemático basado en redes neuronales inspiradas en el comportamiento biológico neuronal de los animales, lo cual permitió dar las bases para los futuros trabajos en redes neuronales artificiales.

Las redes neuronales se encuentran compuestas de una serie de perceptrones los cuales realizan cálculos matemáticos de acuerdo a señales recibidas de los perceptrones anteriores, estos perceptrones o neuronas están dados por la función 1:

$$Y = \Theta \left\{ \sum_{j=1}^{n} w_j x_j - u \right\} \tag{2.1}$$

Donde $\Theta$ es una función de activación, $w_j$ es el peso asociada a la neurona $x_j$ y un umbral $u$, esta función logra modelar el comportamiento de una neurona:

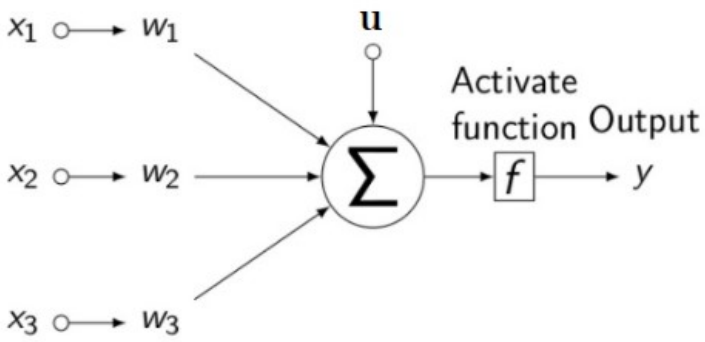


Figure 2.1: Modelo neuronal de McCulloch y Pitts, 1943.

Este modelo neuronal requiere de una función de activación que permita transformar las señales de entrada a una señal de salida adecuada, existen cuatro tipos de funciones de activación que son comúnmente usadas, entre ellas están ReLu, Sigmoid, piecewise, Gaussian y linear.

Es sabido que las redes neuronales constan de unos pesos los cuales deben ser definidos previamente, debido a la complejidad que pueden adquirir las redes y los patrones a reconocer, se debe realizar por medio de un entrenamiento previo, el cual es realizado mediante optimización de parámetros o backpropagation, al igual que

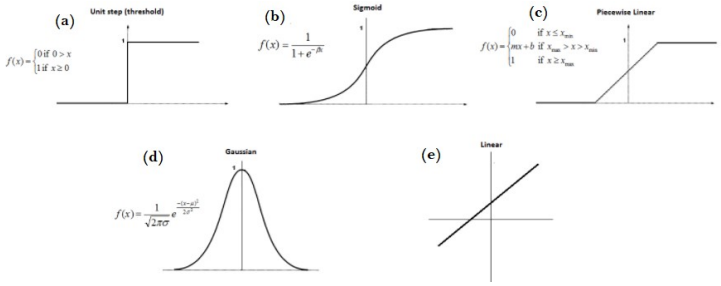


Figure 2.2: Funciones de activación (a) La salida se ajusta en una de las dos condiciones dependiendo de la suma total de la entrada, (b) consta de dos funciones logística y tangencial adquieren valores 0 a 1 y -1 a 1 respectivamente, (c) adquiere un valor proporcional dependiendo de la ponderación total, (d) se interpreta la pertenencia a la clase (1/0) de acuerdo a un valor promedio.

las funciones de activación existe una gran cantidad de algoritmos de aprendizaje (optimización) que permiten ir ajustando los pesos de las conexiones, algunas de estas son: gradiente descendente, gradiente estocástico descendente, Método de Newton, Método quasi-Newton, Levenberg-Marquardt, etc. De los anteriores métodos se encontró que se recomienda el uso de gradiente estocástico descendente LeCun et al. (1998), puesto que optimiza con un tiempo de O(N) para redes profundas y permite realizar una actualización de parámetros por bloques lo cual permite una fácil aplicación en métodos de Deep Learning, este método de gradiente descendente estocástico está definido por la función 2.

$$w_{t+1} = w_t - \gamma_t \bigtriangledown_w Q\left(z_t, w_t\right) \tag{2.2}$$

Cuando $\gamma$ es suficientemente pequeño, se dice que se logra una convergencia lineal, por tanto, en cada iteración se estima el gradiente sobre el bloque de datos escogidos, llevando a la definición de parámetros por medio de optimización.

Hoy en día se desarrollan nuevas ramas de la computación basadas en redes neuronales como ANN convolucionales, ANN recurrentes, ANN Residuales, Deep NN y Deep Belief NN, entre otros.

En esta revisión se han encontrado algunas aplicaciones enfocadas a la clasificación de granos en donde se aplican ANN Multicapa-perceptrón donde la capa de entrada contiene el número de neuronas correspondiente a los descriptores de entrada, la capa de salida contiene el número de neuronas correspondiente a las clases de clasificación, las capas ocultas son propuestas por cada investigación realizada 12, 30, 20, 13 neuronas ( Kılıç et al. (2007), Szczypiński et al. (2015), Kurtulmus et al. (2016), Omid et al. (2017) respectivamente) de acuerdo a pruebas de topología las cuales consisten en ir incrementando la cantidad de neuronas hasta identificar en qué momento se obtiene una mayor precisión, estos modelos de clasificación demuestran precisiones por más del 80% y en un caso llega al 99% Omid et al. (2017).

Kivanc Kilic et al en Kılıç et al. (2007) desarrollaron un sistema de visión por computador para evaluar la calidad de frijoles utilizando ANN, el conjunto de características usado para aplicar en esta técnica fue obtenido por medio de aplicar los momentos (promedio, varianza, skewness y kurtosis) a los tres canales RGB, lo cual permitió cuantificar el color con un 90.6% de precisión. Por otro lado, se encontró en Szczypiński et al. (2015) y Kurtulmus et al. (2016) que las semillas de cebada y pimiento son difíciles de clasificar visualmente aun para expertos, por lo cual se proponen técnicas de automatización donde se realizan pruebas con ANN, para clasificar 11 variedades de cebada y 8 de pimiento, donde se propone obtener un gran conjunto de características (290 y 257 obtenidas de RGB, HSB, YUV, YIQ, CIE XYZ y CIE Lab) el cual es reducido para generar características representativas de las semillas a clasificar (7 y 10 respectivamente), se realiza la aplicación de redes neuronales con 20 y 30 neuronas donde se dan muy buenos resultados, teniendo en cuenta la complejidad de la clasificación de los cultivos.

### 2.2.2. Support vector machine(SVM)

Las SVM mencionadas inicialmente por Boser, B et al Boser et al. (1992) como un algoritmo que maximiza la margen entre los patrones de entrenamiento y los umbrales de decisión, dando paso al planteamiento formal de las Maquinas de Soporte Vectorial las cuales hoy en día presentan 4 tipos (the Maximal Margin Classifier, The kernelized, The soft-margin, The soft-margin kernelized) de los cuales el

más usado es el The soft-margin kernelized que combina las tres técnicas iniciales.

Esta técnica consiste en realizar clasificación binaria encontrando la mayor distancia entre dos clases, aunque es extensible a modelos que requieren tratar problemas multi-clases, estos se deben reducir a problemas binarios utilizando técnicas como uno contra todos (one against all) o uno contra uno (one against one). Estos problemas binarios deben ser linealmente separables y si no es posible realizar esta separación de clases linealmente, se debe aplicar funciones de kernel (RBF) que permitan obtener un hiperplano donde las clases sean linealmente separables.

Inicialmente los modelos de clasificación de SVM son entrenados previamente por un conjunto de datos etiquetados y posteriormente se crea el hiperplano de máximo margen de separación el cual está dado por la función de decisión generalizada:

$$y_i = (w \cdot x) + b \quad \text{donde} \quad w, x \in R^n, b \in R) \tag{2.3}$$

Donde $w$ y $b$ son coeficientes reales, esta función es una restricción base que permite que se cumpla la separación para los valores de $x$, restringe $y_i \geq +1$ Y $y_i \leq -1$.

Posteriormente es necesario definir la separación optima del hiperplano la cual está dada por la expresión.

$$\tau = 1/||w|| \tag{2.4}$$

Esta expresión da el mayor margen de separación para los valores en que las restricciones propuestas en la ecuación 2.3 se presentan como una igualdad, lo que indica que estos valores son los que se consideraran como vectores de soporte.

Las aplicaciones para problemas linealmente no separables ocurren en ocasiones cuando los datos presentan gran variabilidad o en problemas multiclase, se han encontrado aplicaciones de SVM en la clasificación de granos en los que se debe separar más de dos clases o que estas no son linealmente separables como pistacho Omid et al. (2017), cereal Brueckner et al. (2009) y maíz Kiratiratanapruk and Sinthupinyo (2011) donde se demuestran clasificaciones con una precisión de más del 90%, estos modelos fueron configurados obteniendo grandes conjuntos de características por medio de los distintos canales de las imágenes, y en Kiratiratanapruk and Sinthupinyo (2011) se adiciona características de textura basado en la matrix de co-ocurrencias de niveles de gris, estas características en las investigaciones mencionadas anteriormente fueron reducidas a menos de 11, al igual que se aplicó SVM usando Radial Basis Function kernel (RBF) para lograr la separación lineal de las clases. Los parámetros de configuración para estas SVM han sido definidos haciendo uso de cross-validation o por medio de entrenamiento y error para garantizar la precisión del modelo.

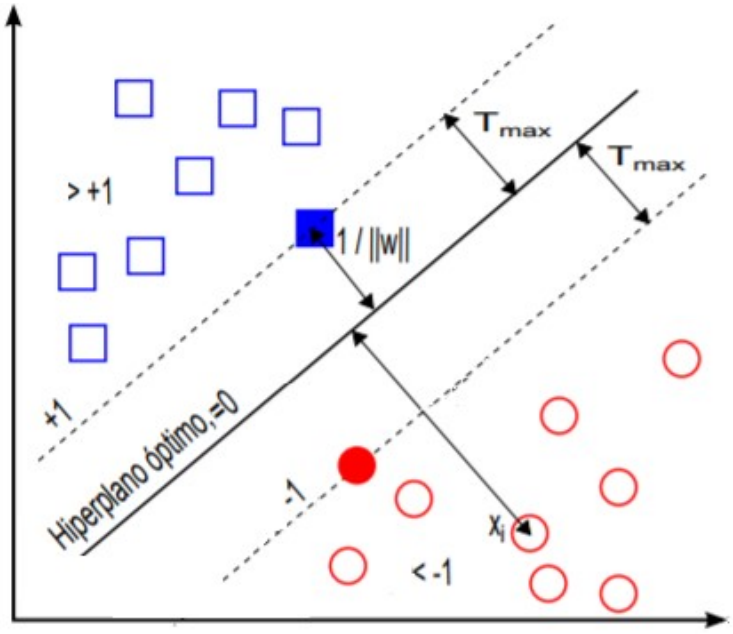


Figure 2.3: Aplicación base conceptos SVM

### 2.2.3. K-Nearest Neighbors(KNN)

Los modelos de clasificación basados en KNN (K-Nearest Neighbors algorithm) Sutton (2012) presenta uno de los enfoques más sencillos de clasificación, esta técnica consiste en identificar los vecinos más cercanos a la nueva muestra a clasificar, basándose en distancias métricas.

(a) $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$ Euclidean

(b) $\sum_{i=1}^{k}|x_i - y_i|$ Manhattan

(c) $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$ Minkowski

(d) $D_H = \sum_{i=1}^{k}|x_i - y_i|$ Hamming

Figure 2.4: Tipos de ecuaciones para determinar las k distancias en un conjunto de entrenamiento, (a) la más comúnmente usada para medir la distancia entre dos puntos, (b) Calcula la distancia entre vectores reales usando la suma de la diferencia en valor absoluto, (c) Es la generalización de la distancia de manhattan y la euclidiana y(d) Usada para calcular la distancia entre vectores binarios.

Este algoritmo recibe un conjunto de datos de entrenamiento etiquetado por medio del cual se definen las clases a clasificar, el algoritmo KNN almacenas estos valores los cuales son usados para asignar la clase de las nuevas muestras. Para clasificar una nueva muestra en el plano se calcula el vector de distancias (ver Figura

2.4) de tamaño $k$ (vecinos) elementos de todos los objetos clasificados previamente, posteriormente se asigna el nuevo punto a la clase que más elementos contenga en el vector de distancias calculado previamente Figura 2.5.

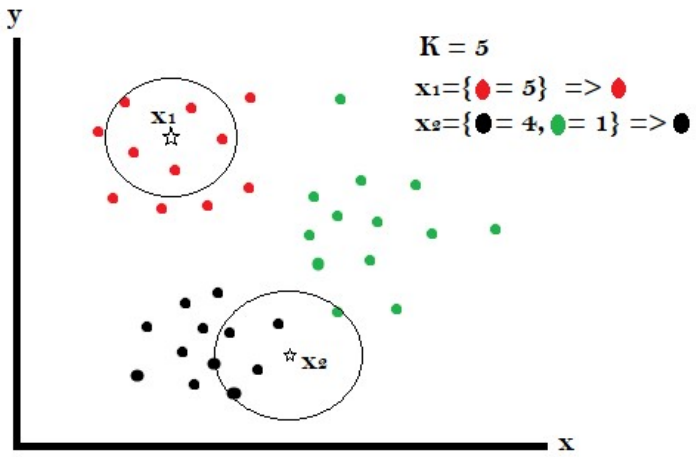


Figure 2.5: Algoritmo K-NN para $k = 5$, clasificación de dos muestras $X_1$ y $X_2$.

Este algoritmo es multiclase pero carece de rendimiento puesto que debe calcular el valor de las distancias de todos los elementos del vector cada vez que ingresa al modelo una muestra nueva a clasificar, también es muy simple puesto que al asignar la misma importancia a todos los valores que aportan en el cálculo de la distancia pierde capacidad de agregar conocimiento de interés en la clasificación, por tanto existen variaciones que permiten mejorar este algoritmo como KNN con rechazo, KNN con distancia media, KNN con distancia mínima, KNN con pesos, KNN con pesos por casos. Para la clasificación de granos de arroz Hansen et al. (2016) para el banco de recursos genéticos del IRRI se realizó la adquisición de imágenes multiespectrales por medio de las cuales se obtuvo un conjunto de 177 características las cuales fueron reducidas a 40 características de importancia, con las cuales se aplica el algoritmo KNN evaluando K desde 1 hasta 30, determinando un K igual a 6 obteniendo una precisión del 93%. Este proyecto menciona la gran importancia que tiene las imágenes multiespectrales para identificar los fenotipos de las semillas puesto que permitió conservar un conjunto de características de importancia mucho mayor a las vistas anteriormente. En Delwiche et al. (2013) se evalúan semillas de trigo por medio de un sistema de caída libre donde se logra capturar la parte frontal de la semillas y dos imágenes opuestas mediante espejos que reflejan la parte anterior de las semillas, en este proyecto se clasifican los tres tipos de daños que presentan las semillas de trigo, por tanto se definió un sistema de caída libre

el cual añadió más tiempo de procesamiento de la imagen para obtener las características apropiadas para la clasificación, pero debido a que se obtuvo una visión más completa de la semilla, se logra obtener más información que permite llevar a una precisión del 93% aplicando KNN con $K = 25$.

### 2.2.4. Other classifiers

Existen otra gran variedad de técnicas de clasificación como Fuzzy logic, Genetic algorithm, SL, Decision tree, en Jayas and Singh (2012) y Shrestha et al. (2015) se realiza una revisión más amplia de las distintas técnicas de clasificación para alimentos (carnes, plantas, granos, etc); donde se menciona la Lógica difusa (Fuzzy logic) como una herramienta que puede tratar reglas generalizadas y obtener resultados complejos cercanos a las decisiones humanas. Los algoritmos genéticos clasificados en ambos artículos como complejos y con una estructura difícil de interpretar que se basan en la búsqueda de información en poblaciones para identificar hipótesis. Técnicas de Statistical Learning (SL) que incluyen aprendizaje bayesiano y análisis discriminante que permiten calcular las distribuciones de probabilidad y conocimiento extraíble u observable de los datos para lograr resultados de clasificación óptimos. Las aplicaciones de estas técnicas han sido pocas o no son muy actuales que sean de interés para el objetivo de esta revisión.

# Chapter 3

# Methodology

## 3.1. Activities

**Activity 1: Análisis.**

- Investigación de modelos de clasificación de imágenes.
- Estandarización del protocolo de captura de imágenes de forrajes.
- Obtención de imágenes y posteriormente aumentar datos.
- Identificación de características fenotípicas y morfológicas de semillas de buena calidad y de mala calidad.
- Selección del algoritmo de clasificación que presente mejores características en cuanto a rendimiento, precisión y tiempo, de acuerdo a sus ventajas y desventajas en la clasificación de semillas.
- Creación de historias de clasificación que realizan los usuarios manualmente.

**Activity 2: Diseño.**

- Diseñar modelo de clasificación binaria de acuerdo a algoritmo de clasificación seleccionado.
- Definir proceso de limpieza y normalización de imágenes.
- Diseñar arquitectura del prototipo para clasificación de imágenes.

**Activity 3: Implementación y evaluación del modelo desarrollado**

- Código fuente.
- Diseño de pruebas de evaluación.
- Evaluación de la precisión de clasificación de acuerdo a la realizada inicialmente por el personal profesional.
- Documento final de evaluación de resultados del proyecto de maestria.

## 3.2. schedule of activities

| Activity | Agosto - Diciembre 2018 — Enero - Julio 2019 | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 08 | 09 | 10 | 11 | 12 | 01 | 02 | 03 | 04 | 05 | 06 | 07 |
| Analysis | X | X | X | X | X | | | | | | | |
| Design | | | | X | X | X | X | X | | X | | |
| Implementation and evaluation | | | | | | | | X | X | X | X | X |

Table 3.1: Cronograma de actividades previstas

## 3.3. Expected result

Se entregara el prototipo junto con la documentacion respectiva a la arquitectura, diseño y pruebas.

# Bibliography


Basheer, I. and Hajmeer, M.: 2000, Artificial neural networks: fundamentals, computing, design, and application, *Journal of microbiological methods* **43**(1), 3–31.

Beyer, K., Goldstein, J., Ramakrishnan, R. and Shaft, U.: 1999, When is "nearest neighbor" meaningful?, *International conference on database theory*, Springer, pp. 217–235.

Boser, B. E., Guyon, I. M. and Vapnik, V. N.: 1992, A training algorithm for optimal margin classifiers, *Proceedings of the fifth annual workshop on Computational learning theory*, ACM, pp. 144–152.

Brueckner, P., Anding, K., Weissensee, H. and Dambon, M.: 2009, Op6-quality assurance of grain with colour line scan cameras, *Proceedings OPTO 2009 & IRS$^2$ 2009* pp. 139–144.

*Conservación y uso de cultivos*: n.d.

Cuervo, M., Martínez, A., Muñoz, L., Ramírez, J., Martínez, J. and Debouck, D. G.: 2016, Manual de procedimientos del laboratorio de sanidad de germoplasma. certificación sanitaria del germoplasma de yuca.

Delwiche, S. R., Yang, I.-C. and Graybosch, R. A.: 2013, Multiple view image analysis of freefalling us wheat grains for damage assessment, *Computers and electronics in agriculture* **98**, 62–73.

Du, C.-J. and Sun, D.-W.: 2006, Learning techniques used in computer vision for food quality evaluation: a review, *Journal of food engineering* **72**(1), 39–55.

*Fenotipos - Fenotipo .com*: n.d.

Grinblat, G. L., Uzal, L. C., Larese, M. G. and Granitto, P. M.: 2016, Deep learning for plant identification using vein morphological patterns, *Computers and Electronics in Agriculture* **127**, 418–424.

Hansen, M. A. E., Hay, F. R. and Carstensen, J. M.: 2016, A virtual seed file: the use of multispectral image analysis in the management of genebank seed accessions, *Plant Genetic Resources* **14**(3), 238–241.

Hernández-Villareal, A.: 2013, Caracterización morfológica de recursos fitogenéticos, *Revista Bio Ciencias* **2**(3).

Jain, A. K., Mao, J. and Mohiuddin, K. M.: 1996, Artificial neural networks: A tutorial, *Computer* **29**(3), 31–44.

Jayas, D. and Singh, C.: 2012, Grain quality evaluation by computer vision, *Computer vision technology in the food and beverage industries*, Elsevier, pp. 400–421.

Kılıç, K., Boyacı, İ. H., Köksel, H. and Küsmenoğlu, İ.: 2007, A classification system for beans using computer vision system and artificial neural networks, *Journal of Food Engineering* **78**(3), 897–904.

Kiratiratanapruk, K. and Sinthupinyo, W.: 2011, Color and texture for corn seed classification by machine vision, *Intelligent Signal Processing and Communications Systems (ISPACS), 2011 International Symposium on*, IEEE, pp. 1–5.

Kurtulmus, F., Alibas, I. and Kavdir, I.: 2016, Classification of pepper seeds using machine vision based on neural network, *International Journal of Agricultural and Biological Engineering* **9**(1), 51.

LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P.: 1998, Gradient-based learning applied to document recognition, *Proceedings of the IEEE* **86**(11), 2278–2324.

McCulloch, W. S. and Pitts, W.: 1943, A logical calculus of the ideas immanent in nervous activity, *The bulletin of mathematical biophysics* **5**(4), 115–133.

Mika, S., Ratsch, G., Weston, J., Scholkopf, B. and Mullers, K.-R.: 1999, Fisher discriminant analysis with kernels, *Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop.*, IEEE, pp. 41–48.

Omid, M., Firouz, M. S., Nouri-Ahmadabadi, H. and Mohtasebi, S. S.: 2017, Classification of peeled pistachio kernels using computer vision and color features, *Engineering in Agriculture, Environment and Food* .

*Plant Production and Protection Division: Las Normas para Bancos de Germoplasma de Recursos Fitogenéticos para la Alimentación y la Agricultura*: n.d.

Shalev-Shwartz, S. and Ben-David, S.: 2014, *Understanding machine learning: From theory to algorithms*, Cambridge university press.

Shrestha, S., Deleuran, L. C., Olesen, M. H. and Gislum, R.: 2015, Use of multispectral imaging in varietal identification of tomato, *Sensors* **15**(2), 4496–4512.

Sutton, O.: 2012, Introduction to k nearest neighbour classification and condensed nearest neighbour data reduction, *University lectures, University of Leicester* .

Szczypiński, P. M., Klepaczko, A. and Zapotoczny, P.: 2015, Identifying barley varieties by computer vision, *Computers and Electronics in Agriculture* **110**, 1–8.

Vithu, P. and Moses, J.: 2016, Machine vision system for food grain quality evaluation: A review, *Trends in Food Science & Technology* **56**, 13–20.

Wang, L.: 2005, *Support vector machines: theory and applications*, Vol. 177, Springer Science & Business Media.

Zhang, H.: 2004, The optimality of naive bayes, *AA* **1**(2), 3.