PROYECTO DE INVESTIGACIÓN

Proyecto Interno ❑ Proyecto Semilla ❑ Proyecto Junior ■ Proyecto Multi e Inter Disciplinario ❑

Investigación Básica ❑ Investigación Aplicada ■ Investigación Pedagógica ❑ Innovación ❑

DEPARTAMENTO(S):
1. Centro de Modelización Matemática: ModeMat
2. Departamento de Matemática

LINEA(S) DE INVESTIGACIÓN:
1. Optimización matemática y control
2. Modelización matemática y cálculo científico

1 Proyecto de Investigación

Título: RECONOCIMIENTO Y CONTEO AUTOMÁTICO DE VECTORES TRANSMISORES DE ENFERMEDADES INFECCIOSAS MEDIANTE APRENDIZAJE DE MÁQUINAS

Resumen del proyecto (máximo 200 palabras)

Se conoce que los agentes responsables de la transmisión de enfermedades como el Dengue, Malaria, Chikungunya, entre otros, son principalmente las especies de mosquito: Aedes–Aegypti, Anofeles y Cúlex; conocidos en epidemiología como *vectores*. En este proyecto planteamos el desarrollo de una metodología para el reconocimiento y conteo automático de diferentes estadios de maduración de vectores asociados a la transmisión de enfermedades mediante fotografías digitales. Para este fin, se propone el desarrollo de modelos matemáticos de la teoría de aprendizaje de máquinas y visión computacional, así como algoritmos para su resolución numérica, los que serán implementados en una herramienta computacional para el procesamiento eficiente de las muestras fotografiadas por investigadores del área de epidemiología del Instituto Nacional de Investigación Pública (INSPI).

Palabras clave (4-6):

Reconocimiento automático de vectores, Problemas de Optimización, Aprendizaje de Máquinas, Métodos Numéricos.

5 Objetivos, relevancia, productos y resultados esperados de esta propuesta de investigación

5.1 Objetivos

5.1.1 Objetivo General

- Desarrollar una metodología basada en modelos matemáticos de aprendizaje de máquinas e implementar una herramienta informática para el reconocimiento y conteo automático de vectores en sus diferentes etapas de maduración: huevos, larvas y pupas.

5.1.2 Objetivos Específicos

Como objetivos específicos de este proyecto tenemos:

a. Establecer un flujo de trabajo para el filtrado y procesamiento de las muestras fotografiadas.
b. Desarrollar un modelo de optimización de máquinas de soporte vectoriales (SVM) para clasificación de imágenes de muestras de huevos, larvas y pupas de los vectores asociados con la transmisión de enfermedades infecciosas en zonas vulnerables del Ecuador.
c. Desarrollar algoritmos numéricos eficientes para la resolución de los modelos de clasificación de SVM.
d. Implementar los algoritmos numéricos para el procesamiento de imágenes de muestras recolectadas.
e. Desarrollar una metodología de preprocesamiento de muestras para el filtrado y segmentación de objetos de una imagen y su posterior identificación y clasificación.

5.2 Relevancia de esta propuesta de investigación y su relación con la(s) Línea(s) de investigación asociadas.

El proyecto es de relevancia en el área de optimización matemática, modelización matemática y cálculo científico en el área de la salud y cumple con los objetivos y líneas de investigación del Departamento de Matemática y del Centro de Modelización Matemática: ModeMat.

5.3 Productos esperados

a. Publicaciones científicas (obligatorio); ■
b. Disertación a la Comunidad Politécnica; ■
c. Proyecto de Titulación; ❑
d. Tesis de Grado (maestría o doctorado); ■
e. Aplicación tecnológica construida o implementada; ❑
f. Patente presentada; ❑
g. Perfil de proyecto de mayor impacto científico, técnico, pedagógico o de innovación. ❑

5.4 Detalle de los resultados esperados (con relación a los objetivos**)**
Al finalizar este proyecto se espera contar con los siguientes resultados y productos:

a. Una metodología que permita el reconocimiento y clasificación automática de las diferentes etapas de maduración de vectores en fotografías digitales de las muestras utilizadas por los investigadores del INSPI, para facilitar el procesamiento de la información entomológica del mosquito.
b. Una herramienta computacional que implemente los modelos y algoritmos desarrollados en la metodología propuesta para el reconocimiento y conteo automático de vectores.

6 Descripción, metodología y cronograma de trabajo

En epidemiología, es conocido que el Aedes-Aegypti es el principal vector transmisor del virus del Dengue en todo el mundo [14] y, más recientemente, se ha asociado esta especie de mosquito a la transmisión de la Chikungunya en el Ecuador; mientras que el Anofeles está asociado con las Transmisión de la Malaria cf. [14]. Estos vectores representan un grave problema en la administración de la salud pública, por lo cual son objeto de estudio en instituciones del área salud pública como el INSPI. Una de las dificultades en la lucha contra la proliferación de epidemias transmitidas por estos vectores, es la gran cantidad de tiempo invertida en procesar las muestras tomadas por biólogos con el fin de recolectar información.

Las técnicas de *aprendizaje de máquinas* han sido exitosamente aplicadas en un sin número de problemas de clasificación y predicción, basándose en datos que contienen información sobre un problema específico, como por ejemplo: clasificación de correo basura (*spam*), reconocimiento de escritura a mano, predicción de cáncer, entre una vasta cantidad de aplicaciones cf.[1]. Bajo ciertas hipótesis (cf.[4],[8]) casi cualquier fenómeno, del cual se disponga adecuada y suficiente información, puede ser sujeto a la clasificación y predicción de sus variables que están altamente correlacionadas con los datos disponibles. El reconocimiento de objetos en imágenes ha sido ampliamente estudiado con diferentes técnicas de aprendizaje de máquinas conjuntamente con la teoría de *visión computacional*, por ejemplo para el reconocimiento facial, guía de vehículos motorizados, etc. (ver por ejemplo [1], [13] o [4]).

Sin embargo, existe poco desarrollo de este tipo de herramientas para el reconocimiento de vectores transmisores de enfermedades como el Aedes-Aegytpi y otros vectores similares. Algunas técnicas han sido desarrolladas con el fin de dar soporte a la tarea de procesar las muestras de manera automática; por ejemplo, en [9] se desarrolla un método para el conteo automático de huevos depositados por el Aedes-Aegypti en trampas diseñadas para mosquitos, que utilizan principalmente herramientas de procesamiento de imágenes para el conteo de los huevos. Sin embargo, el conteo se basa más bien en técnicas estadísticas y la técnica se limita solamente al conteo de huevos. En [12] se reporta el desarrollo de un paquete computacional para la clasificación de larvas, para la diferenciación entre las subespecios Aedes-Aegypti y Aedes-albopictus utilizando imágenes de alta resolución. Por tanto, el desarrollo de una metodología para el procesamiento de vectores para su conteo y clasificación automática de sus fases de maduración sería un aporte muy importante para el monitoreo de vectores que realiza el INSPI.

EL proceso de detección de una muestra empieza capturando una imagen de la muestra en alta resolución. La fotografía captura una bandeja con agua que contiene los especímenes y otras impurezas en condiciones adecuadas de iluminación. En este proyecto nos concentramos en la clasificación en las etapas de maduración de vectores de Aedes–Aegypti, Cúlex y Anofeles ver [10], [15] . La fotografía es procesada y la cantidad de larvas, huevos y pupas es contabilizada sin diferenciar la especie, de acuerdo al siguiente flujo de trabajo:

1. Preparación de la muestra: Las muestras serán tratadas en contenedores estándar en los laboratorios del INSPI a fin de poseer una cantidad de larvas y pupas en un volumen constante para todas las muestras volumen no muy elevado elevado que permita el movimiento u ocultamiento de alguna larva al lente de la cámara. Finalmente, se procederá a fotografiar la bandeja con las larvas para su futura cuantificación larval, diferenciándolas por estadios.
2. Preprocesamiento: Debido a que la muestra puede contener impurezas como es necesario eliminar el ruido inducido por las impurezas en la imagen. Para esto, se aplicará la técnica de eliminación de ruido (*denoising*) aplicada en [11], para lo cual se planteará un problema de optimización con un funcional de variación total; es decir, nos interesa recuperar una imagen sin ruido. En este modelo se minimiza gradiente de la imagen, considerada como una función bidimensianal, para garantizar la suavidad de la función y evitar así posibles efectos “blur” en la imagen original. Posteriormente, se procede a segmentar la imagen para el reconocimiento de las formas, para así delimitar los especímenes de vectores en regiones de la imagen que serán sometidos al análisis de clasificación. En este proceso es posible que se segmenten impurezas de tamaño más grande o insectos de otras especies que no fueron eliminados mediante el filtro de ruido que habrá que tomar en cuenta en fases posteriores y evitar falsos positivos en el proceso de reconocimiento. La segmentación será realizada a través de un modelo de Chan-Vese cf.[2], utilizando funciones de nivel.
3. Fase de clasificación: Una vez que los potenciales especímenes han sido localizados se procede a utilizar un modelo basado en máquinas de soporte vectoriales (SVM) para la clasificación en larvas, pupas o huevos que hayan sido seleccionadas en la fase de segmentación; que además discriminará objetos que no correspondan a vectores. Para el diseño de la SVM se seguirán los siguientes pasos:

4.1. Extracción de características de los especímenes: Se debe construir una base de datos de las características morfológicas y biométricas que diferencien los diferentes estados de maduración de los vectores. Para esta tarea se recolectará una cantidad adecuada de imágenes de características homogéneas (en cuanto a tamaño y resolución) de: huevos, pupas y larvas; de las cuales serán preprocesadas. En lugar de usar las imágenes directamente, la idea es realizar un estudio sobre características (*features*) que describan los patrones claves de la estructura morfológica de los vectores; como por ejemplo utilizando histogramas de gradientes, en forma análoga a los utilizados en [7] y [3], o como características de tipo Haar cf. [6].

4.2. Análisis de correlación: Se realizará un análisis de componentes principales (cf. [4]) con el fin de determinar las características más relevantes para el proceso de entrenamiento, necesaria para mantener el tamaño del modelo lo más pequeño posible.

4.3. Entrenamiento: Se formulará un modelo de SVM con Kernel de bases de funciones radiales (RBF) y de tipo logístico. La SVM consiste en un problema de optimización para encontrar los coeficientes del clasificador cf. [4],[8]. Se propondrán diferentes técnicas de regularización de la función objetivo: con norma-2 y norma-1 cf. [5] y se realizará un análisis del problema para establecer las condiciones de optimalidad que caracterizan la solución cf. [13].

4.4. Clasificación: Se determinará una lista de clasificación, que consiste en la lista de “individuos" que se quiere clasificar, ie. los diferentes estados de maduración de los vectores y objetos extraños que no correspondan a especímenes del mosquito. Se construirá un modelo entrenado (SVM) para cada espécimen y se utilizará un esquema *one versus all* para clasificar de acuerdo a la lista de clasificación cf. [4],[8].

4.5. Resolución numérica: Se desarrollarán algoritmos numéricos eficientes para su resolución, basados en algoritmos de primer y segundo orden ver cf. [14] y se realizarán pruebas de validación para calibrar parámetros de los modelos y los algoritmos para obtener los resultados deseados, así como prevenir efectos de “overfitting” para lograr un alto poder de clasificación. Finalmente, se realizará una validación comparando con los resultados de los algoritmos para la verificar la efectividad del reconocimiento.

Bibliografía

[1] Christopher M Bishop.(2006). *Pattern recognition and machine learning*. Springer.

[2] Ethan S Brown, Tony F Chan, and Xavier Bresson. (2012). Completely convex formulation of the chan-vese image segmentation model. *International journal of computer vision*, 98(1):103–121.

[3] Navneet Dalal and Bill Triggs. (2015). Histograms of oriented gradients for human detection. volume 1, pages 886–893. IEEE.

[4] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin.

[5] Haoying Fu, Michael K Ng, Mila Nikolova, and Jesse L Barlow. (2006). Efficient minimization methods of mixed l2-l1 and l1-l1 norms for image restoration. *SIAM Journal on Scientific computing*, 27(6):1881–1902.

[6] Natalia Larios, Bilge Soran, Linda G Shapiro, Gonzalo Martìnez-Muñoz, Junyuan Lin, and Thomas G Dietterich. (2006). Haar random forest features and svm spatial matching kernel for stonefly species identification. pages 2624–2627. IEEE.

[7] David G Lowe. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110.

[8] Stephen Marsland. (2014). *Machine learning: an algorithmic perspective*. CRC press.

[9] Carlos AB Mello, Wellington P dos Santos, Marco AB Rodrigues, Ana Lucia B Candeias, and Cristine MG Gusmão. (2008). Image segmentation of ovitraps for automatic counting of aedes aegypti eggs. pages 3103–3106. IEEE.

[10] Yasmin Rubio-Palis, Escuela de Malariologia, Saneamiento Ambiental, and Arnoldo Gabaldón. (1998). Caracterización morfométrica de poblaciones del vector de malaria anopheles (nyssorhynchus) darlingi root (diptera: Culicidae) en venezuela. *Boletin de Entomologia Venezolana*, 13(2):141–172.

[11] Leonid I Rudin, Stanley Osher, and Emad Fatemi. (1992) Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268.

[12] André Iwersen de São Thiago, Emil Kupek, Joaquim Alves Ferreira Neto, and Paulo de Tarso São Thiago. (2002) Software for pattern recognition of the larvae of aedes aegypti and aedes albopictus. *Revista da Sociedade Brasileira de Medicina Tropical*, 35(3):263–265.

[13] Suvrit Sra, Sebastian Nowozin, and Stephen J Wright. (2012). *Optimization for machine learning*. Mit Press.

[14] Willem Takken and Bart GJ Knols. (2007) *Ecology and control of vector-borne diseases*. Wageningen Acad. Publ.

[15] LARS ERIK Widahl. (1988). Some morphometric differences between container and pool breeding culicidae. *Journal of the American Mosquito Control Association*, 4(1):76–81.

6.2 Cronograma de trabajo anual: *(Descripción)*

- *Para la elaboración del cronograma de ejecución del proyecto se sugiere considerar el tiempo para la adquisición de equipos, reactivos y materiales de laboratorio.*

Primer Año

	Porcentaje de avance por mes						TOTAL
Actividad	**1-2**	**3-4**	**5-6**	**7-8**	**9-10**	**11-12**	
Compra de Equipos y bibliografía	**3%**	**2%**					**5%**
Contratación de investigadores y pasantes		3%	2%				5%
Toma de fotografías			5%	5%	5%		15%
Desarrollo de la metodología de preprocesamiento: filtrado, eliminación de ruido y segmentación		5%	5%	5%	10%	10%	35%
Extracción de características de las muestras y análisis				15%	15%	10%	40%
TOTAL	3%	10%	12%	25%	30%	20%	100%

Segundo Año 2

	Porcentaje de avance por mes						TOTAL
Actividad	**1-2**	**3-4**	**5-6**	**7-8**	**9-10**	**11-12**	
Contratación de investigadores y pasantes	**3%**	**2%**					5%
Diseño de la SVM		15%	10%	10%			35%
Implementación del algoritmo de optimización			15%	10%	10%		35%
Test numéricos y validación					10%	15%	25%
TOTAL	3%	17%	25%	20%	20%	15%	100%

7 Fechas de inicio y fin

Fecha de inicio: 1 de enero de 2016
Fecha de finalización:31 de diciembre de 2018

8 Tiempo de dedicación de docentes, infraestructura, equipos y fondos adicionales.

8.1 Tiempo máximo de dedicación semestral del Director del proyecto, de los docentes participantes y otros colaboradores.

Director: 20 Horas semanales
Colaborador: 10 Horas semanales

8.2 Infraestructura y equipos
El laboratorio de cálculo científico del centro de modelizacón cuenta con un HPC (supercomputador) que tiene capacidad de expansión y cuenta con un datacenter con capacidades de climatización para albergar para un servidor que tenga capacidad de cálculo con GPU y se integre a la plataforma existente. Cabe señala, que actualmente el HPC no cuenta con equipos de procesamiento en GPU.

8.3 Breve justificación del equipo requerido
Debido a que en este proyecto tendrá una carga pesada con trabajo de imágenes de alta resolución se requiere dos computadores de escritorio con monitores de alta resolución y alta capacidad gráfica.
Para el proceso de resolución numérica y cálculo numérico se requiere un servidor con capacidades de cálculo en GPU, necesarias para el pre-procesamiento (segmentación, eliminación de ruido, operaciones y filtros gráficos) de las imágenes, así como su tratamiento numérico y la resolución de los modelos de optimización propuestos.

8.4 Fondos Adicionales
- *No existen fondos adicionales.*