Sistema de clasificación para predicción de fracasos en implantes dentales validado por expertos humanos

Nancy B. Ganz[1], Alicia E. Ares[1] y Horacio D. Kuna[2]

[1] Instituto de Materiales de Misiones (IMAM-CONICET), Facultad de Ciencias Exactas Químicas y Naturales, Universidad Nacional de Misiones, Posadas, Misiones, Argentina.
[2] Instituto de Investigación, Desarrollo e Innovación en Informática (IIDII), Facultad de Ciencias Exactas Químicas y Naturales, Universidad Nacional de Misiones, Posadas, Misiones, Argentina.
{nancy.bea.ganz, a.e.ares, hdkuna}@gmail.com

Resumen. Hoy en día, la predicción del éxito o fracaso de un implante dental está determinado a través de una evaluación clínica y radiológica. Por esta razón, las predicciones dependen en gran medida de la experiencia del implantólogo. Este trabajo tiene por objetivo investigar el beneficio de la utilización de múltiples algoritmos de clasificación, para la predicción de fracasos en implantes dentales de la provincia de Misiones, Argentina validado por expertos humanos. El modelo abarca la combinación de los clasificadores Random Forest, SVM, KNN, Naive Bayes y perceptrón multicapa. La experimentación es realizada con cuatro conjuntos de datos, un conjunto de implantes dentales confeccionado para el estudio de caso, un conjunto generado artificialmente y otros dos conjuntos obtenidos de distintos repositorios de datos. Nuestro enfoque logra sobre el conjunto de datos de implantes un porcentaje de acierto del 93% de casos correctamente identificados, mientras que los expertos humanos consiguen un 86% de precisión. En base a esto podemos alegar, que los sistemas de múltiple clasificadores son un buen enfoque para la predicción de fracasos en implantes dentales.

Palabras claves: ensamble de clasificadores, predicción, clasificación, fracaso, implantes dentales.

1 Introducción

La integración de clasificadores puede ser elemental a la hora de tomar decisiones, debido a que trata de obtener la solución más eficiente para un problema en cuestión. Es posible integrar decisiones obtenidas con el mismo o distintos clasificadores de base [1]. La integración suele ser más precisa, porque los datos de entrenamiento pueden no proporcionar información suficiente para elegir el mejor clasificador y en esta situación la combinación es la mejor opción [2].

En este trabajo se estudia la aplicación de varios métodos de clasificación para la predicción del resultado postoperatorio (éxito o fracaso) de un conjunto de datos de implantes dentales. Con el objetivo de aumentar el acierto de los fracasos. El procedimiento que se propone utiliza los clasificadores: Random Forest (RF) [3], SVM

[4], KNN [5], Naive Bayes (NB) [6] y perceptrón multicapa (MLP) [7]. El ensamble consistió en aplicar pesos a los clasificadores y promediar sus predicciones.

La contribución de este trabajo es un enfoque de aprendizaje automático para la predicción en implantes dentales, el cual es un dominio de poco conocimiento. Asimismo, demostramos que los sistemas de múltiples clasificadores también pueden ser aplicados al estudio de caso, ya que permite lograr mejores rendimientos de clasificación que los alcanzados de manera individual por los clasificadores.

2 Trabajos Relacionados

Existen diversos trabajos de investigación sobre la combinación o integración de clasificadores, para mejorar el acierto de predicción o inclusive para no sesgar la decisión sobre los resultados de un solo clasificador [8]. A continuación, se presentan y examinan algunos trabajos previos sobre este tema.

Miao et al. [9] proponen un procedimiento para mejorar la precisión en la identificación de genes mediante la integración de los clasificadores SVM, RF y SVM. Luego del entrenamiento y predicción con los tres clasificadores, los resultados fueron combinados a través del método de votación mayoritaria [8]. Lograron obtener una precisión mayor a través de la integración de las predicciones que de forma individual. De igual manera, Catal y Nengir [10] han utilizado los clasificadores NB y SVM mediante la técnica de voto mayoritario para la integración de las predicciones. Demostraron a través de sus experimentaciones sobre varios conjuntos de datos que los sistemas de clasificadores múltiples mejoran la precisión. Otro trabajo de similares características es el de Pandey y Taruna [11], donde proponen un clasificador integrado utilizando un árbol de decisión J48, KNN y agregación de estimadores de una dependencia, sobre un conjunto de datos de rendimiento académico de estudiantes de ingeniería. En este modelo, cada clasificador individual genera su valor de predicción y se integran a través del producto de las probabilidades, donde la etiqueta de clase final está representada por el máximo de una probabilidad posterior.

Inspirados en las ideas anteriores, proponemos un procedimiento mediante la utilización de múltiples clasificadores para la predicción del resultado postoperatorio en implantes dentales. El enfoque propuesto fue capaz de superar el porcentaje de acierto de los fracasos logrado por cada clasificador.

3 Materiales y Métodos

3.1 Selección de características

Un paso significativo en el aprendizaje automático es la selección de características, ya que elimina características irrelevantes y redundantes, logrando reducir la dimensionalidad y los requisitos de cálculo, así como puede mejorar el rendimiento de los clasificadores. Su propósito es encontrar un subconjunto óptimo de características que proporcione buenos resultados en la predicción [12], [13]. Se empleo Chi-Square (X^2), el cual es un método muy utilizado para la selección de características categóricas

[14], [15]. Está dado por: $X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$ donde O_{ij} es la frecuencia observada y E_{ij} es la frecuencia esperada. Cuanto mayor sea el valor de X^2, mayor será la evidencia de correlación entre las dos características. El criterio de corte fue la utilización de un nivel de significación. Por lo general los investigadores eligen un nivel de significancia igual a 0.01, 0.05 o 0.10, pero puede ser cualquier valor entre 0 y 1 [15], [16]. Para este trabajo se plantea un nivel de significancia de $p \leq 0.05$ para todos los conjuntos de datos.

3.2 Estructura de los conjuntos de datos

Luego de la selección de características, es necesario dividir los datos. Una estrategia común consiste en tomar todos los datos etiquetados y dividirlos en subconjunto de entrenamiento y evaluación, normalmente con una proporción del 70 al 80 % para entrenamiento y un 20 al 30 % para evaluación o prueba [14], [17], [18]. Esta división va a depender en gran medida del número total de muestras y del modelo a entrenar [19], [20]. En nuestro caso se dividió los datos de forma aleatoria para preservar la distribución de ambas clases en: 70 % para entrenamiento y 30 % para evaluación [21]–[24]. Garantizando que todos los casos se encuentren representados en ambos conjuntos. En la tabla 1 se presentan las características resumidas de los conjuntos de datos utilizados para la experimentación.

Tabla 1. Características de los conjuntos de datos utilizados para la evaluación experimental. De izquierda a derecha se presenta: nombres de los conjuntos de datos, número de muestras, número de atributos por tupla, cantidad de características seleccionadas por el método X^2 y tamaño de los conjuntos de entrenamiento y de prueba.

Conjunto de datos	Muestra	Características	X^2	Entrenamiento	Prueba
Implantes Dentales[1]	1165	33	17	815	350
Artificial[2]	1748	33	21	1223	525
Heart Disease[3]	303	13	10	212	91
Breast Cancer[4]	277	10	5	193	84

[1]***Implantes Dentales:*** conjunto de datos de historias clínicas de pacientes que se han sometido a procesos quirúrgicos de colocación de implantes dentales en la Provincia de Misiones, Argentina. Se encuentra representado a través de 4 dimensiones: datos del paciente (antecedentes y condiciones médicas de los pacientes a la hora de la intervención), datos del implante (características del implante utilizado por el especialista implantólogo), datos de la fase quirúrgica (procedimiento de intervención quirúrgica y mejoramiento del lecho óseo del paciente) y datos del seguimiento postoperatorio (resultado del proceso de colocación del implante, es decir si el proceso de oseointegración implante/tejido-óseo tuvo éxito o fracasó).
[2]***Artificial:*** conjunto artificial generado con el algoritmo SMOTE [25] en base al conjunto *Implantes Dentales*.
[3]***Heart Disease:*** conjunto de datos relacionado a la presencia o ausencia de enfermedad cardíaca en pacientes [26].
[4]***Breast Cancer:*** conjunto de datos relacionado a registros de cáncer de mama que se obtuvieron en el Instituto de Oncología del Centro Médico Universitario de Ljubljana, Yugoslavia [27].

3.3 Proceso de entrenamiento

Para obtener un modelo robusto y optimizar los resultados de los clasificadores, se realizó una búsqueda en cuadrícula para ajustar los híper parámetros [21], [24], [28]. Esta búsqueda se efectuó con los datos de entrenamiento de cada uno de los conjuntos de datos. Para este proceso se especificó:

1. Un espacio de búsqueda, se definió rangos de valores para los híper parámetros y se fue ajustando en función de la medida de rendimiento.
2. Un algoritmo de optimización o ajuste, se empleó el método GridSearchCV [29], es el más costoso en cuanto a rendimiento, pero permite cubrir todo el espacio de búsqueda definido.
3. Un método de evaluación, como estrategia de remuestreo se utilizó validación cruzada de 10 iteraciones.
4. Una medida de rendimiento, se fijó la métrica precisión de equilibrio, la cual está dada por los verdaderos positivos más los verdaderos negativos dividido por la totalidad de muestras del conjunto de datos [30].

En la Tabla 2, se exponen los híper parámetros que se buscó ajustar para lograr el mejor desempeño cada clasificador sobre cada conjunto de datos, además se detallan los espacios de búsquedas definidos para cada parámetro.

Tabla 2. Híper parámetros y espació de búsqueda definido en los clasificadores individuales.

Clasificadores	Parámetros	Espacio de búsqueda
RF	*n_estimators*	*range (1, 150)*
	criterion	*gini, entropy*
SVM	*kernel*	*linear, rbf, poly*
	C	*range (1, 10)*
	gamma	*range (1, 10)*
	degree	*range (1, 10)*
KNN	*n_neighbors*	*range (1, 100)*
	weights	*uniform, distance*
	p	*manhattan, euclidean*
NB	*alpha*	*[0, 0.1, 0.2, 0.3, …, 0.9, 1]*
	fit_prior	*True, False*
MLP	*hidden_layer_sizes*	*range (1,50)*
	activation	*logistic, tanh, relu*
	alpha	*[0.0001, 0.05]*
	solver	*lbfgs, sgd, adam*

3.4 Integración de las predicciones

Para determinar la etiqueta de clase final, en este trabajo se aplica el método de votación suave ponderada [31], [32]. Esta regla permitió lograr los mejores resultados de predicción para el estudio de caso. Por lo tanto, la integración de las predicciones consistió en multiplicar para cada tupla el valor de probabilidad de cada clase, obtenida

por cada clasificador por el peso asignado al mismo. El peso fue determinado mediante una búsqueda en cuadricula utilizando un parámetro de prueba w con valores comprendidos entre 0 y 1. Esta búsqueda fue sometida a una validación cruzada, donde se medió la precisión de cada clasificador para la clase en cuestión, seleccionando el valor de w que logró la mejor precisión [33], [34].

Una vez determinado los pesos, se aplicó el método de votación suave ponderada [31], [32]. Este método recoge las probabilidades de clase predichas por cada clasificador, multiplica por el peso asignado al mismo y los promedia. La etiqueta de clase final se deriva de la etiqueta de clase con la probabilidad promedio más alta. Está dado por: $\hat{y} = \arg max_i \sum_{j=1}^{m} w_j p_{ij}$ donde p_{ij} es la probabilidad predicha por el jth clasificador, w_j es el peso asignado al jth clasificador.

En el presente trabajo en lugar de utilizar el promedio máximo aplicamos un umbral [2], [35], ya que en evaluaciones exploratorias nos permitió lograr mejores resultados en la clasificación. Este umbral estuvo determinado por una búsqueda en cuadricula utilizando un parámetro de prueba μ con valores comprendidos entre 0.1 y 0.5 con incrementos de 0.1 en cada prueba. Se seleccionó el valor de μ que permitió obtener el mejor resultado de clasificación para todos los conjuntos de datos utilizados.

3.5 Parámetros de evaluación

Los parámetros utilizados para evaluar el rendimiento de los clasificadores individuales y comparar con el enfoque propuesto, fueron: matriz de confusión, sensibilidad, especificidad, precisión y error [30], [36].

3.6 Clasificación por expertos humanos

El rendimiento a nivel humano permite estimar una tasa de error óptima y corroborar el funcionamiento del sistema de clasificación. Para evaluar el rendimiento del enfoque propuesto sobre el conjunto de datos de *Implantes Dentales*, se realizó una comparación con la opinión de expertos humanos. Estos fueron seleccionados del "Registro de Profesionales que practican Cirugía Buco maxilofacial, Implantología, Periodoncia y Manipulación de Tejidos del Colegio de Odontólogos de Misiones".

La evaluación estuvo sujeta a la clasificación por cuatro expertos del área, a cada uno de ellos se le suministro una muestra aleatoria distinta del 10% de prevalencia de los casos. Los casos fueron presentados sin la etiqueta para que el experto lo clasifique en función de su experiencia, y de esta manera poder contrastar con los valores hallados por nuestro enfoque de clasificación.

3.7 Enfoque propuesto

En la figura 1 se presenta un diagrama resumen del enfoque propuesto en este trabajo de investigación.

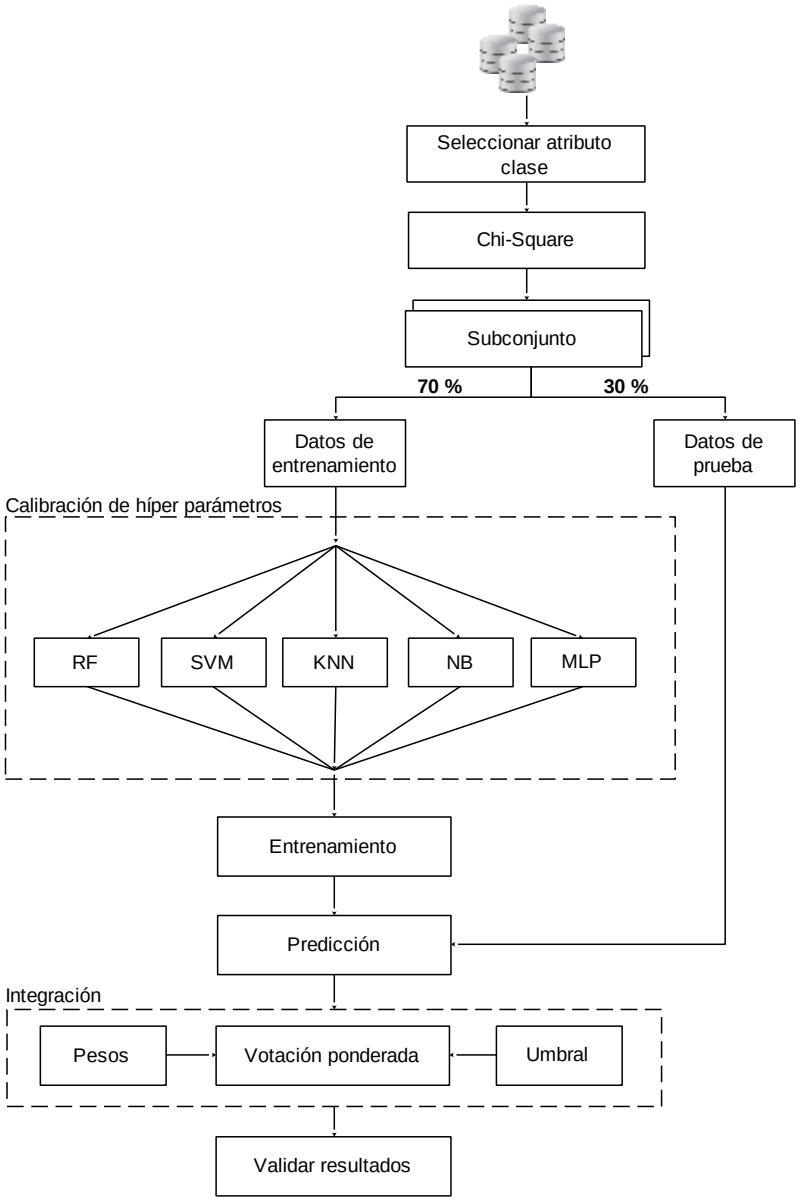

Fig. 1. Enfoque propuesto para la integración de las predicciones de múltiples clasificadores.

4 Resultados

En la tabla 3 se presentan los porcentajes de acierto de las clases objetivo, obtenidas por cada clasificador de forma individual así como del enfoque propuesto sobre los datos de prueba de cada conjunto.

En esta tabla se puede apreciar que el rendimiento de la integración de las predicciones, fue la mejor opción en el acierto de la etiqueta de clase objetivo de todos los conjuntos de datos empleados, ya que logró alcanzar el mayor porcentaje de acierto. Mientras que los clasificadores de forma individual lograron un rendimiento inferior que el de la integración de los mismos.

Tabla 3. Eficiencia en el acierto de la etiqueta de clase objetivo de los clasificadores RF, SVM, KNN, NB, MLP y el enfoque propuesto (Integrado) sobre los conjuntos de datos *Implantes Dentales*, *Artificial*, *Heart Disease* y *Breast Cancer*.

Conjunto de Datos	Clasificadores	% de acierto
Implantes Dentales	RF	59 %
	SVM	64 %
	KNN	64 %
	NB	72 %
	MLP	66 %
	Integrado	**75 %**
Artificial	RF	81 %
	SVM	81 %
	KNN	81 %
	NB	60 %
	MLP	82 %
	Integrado	**89 %**
Heart Disease	RF	81 %
	SVM	70 %
	KNN	70 %
	NB	77 %
	MLP	72 %
	Integrado	**90 %**
Breast Cancer	RF	36 %
	SVM	36 %
	KNN	20 %
	NB	52 %
	MLP	32 %
	Integrado	**58 %**

Finalmente, se comparó los resultados logrados por el enfoque propuesto sobre el conjunto de datos *Implantes Dentales*, con la precisión lograda en la clasificación por los expertos humanos (tabla 4). Nuestro modelo logró un 93% de precisión total, con

un error del 7%. Mientras que en promedio la clasificación realizada por parte de los expertos, logró una precisión total del 86%, con un error promedio del 14%.

Tabla 4. Parámetros de evaluación logrados por el enfoque propuesto y la clasificación de los expertos sobre el conjunto de datos *Implantes Dentales.*

Modelo	Sensibilidad	Especificidad	Precisión	Error
Enfoque propuesto	75 %	96 %	93 %	7 %
Expertos humanos	73 %	92 %	86 %	14 %

5 Conclusiones y futuras líneas de investigación

Este trabajo permitió el estudio de la aplicación de múltiples clasificadores a un dominio de poco conocimiento.

Según los resultados experimentales, el enfoque de múltiple clasificadores también puede ser aplicado a la predicción de fracasos en implantes dentales.

En base a los resultados de la clasificación por parte de los expertos humanos, podemos decir que nuestro enfoque permitió lograr un rendimiento de clasificación superior. Por lo tanto, hemos logrado proponer un procedimiento de extracción de conocimiento validado por expertos humanos.

Finalmente, se plantea como trabajo futuro validar el enfoque propuesto con otros conjuntos de datos del área de la salud o la medicina. Además, se propone la inclusión o ampliación de los clasificadores utilizados, para evaluar la posibilidad de ajustar el porcentaje de acierto de ambas clases. Así como, extender el relevamiento de casos de historias clínicas de implantes dentales al Nordeste argentino, como a otras partes del territorio nacional e internacional.

Agradecimientos
Agradecemos al Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) por el financiamiento a través de una beca doctoral y a los especialistas implantólogos que colaboraron en la construcción de la base de datos de implantes dentales de la presente investigación.

Referencias

[1] Y. Lu, “Knowledge integration in a multiple classifier system,” *Appl. Intell.*, vol. 6, no. 2, pp. 75–86, 1996.

[2] M. Mohandes, M. Deriche, and S. O. Aliyu, “Classifiers Combination Techniques: A Comprehensive Review,” *IEEE Access*, vol. 6, pp. 19626–19639, 2018.

[3] L. Breiman, “Random Forest,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[4] C. Chang and C. Lin, “LIBSVM : A Library for Support Vector Machines,” *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–39, 2011.

[5] N. S. Altman, "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression," *Am. Stat.*, vol. 46, no. 3, pp. 175–185, 1992.

[6] C. D. Manning, P. Raghavan, and H. Schutze, "Text classification and Naive Bayes," in *Introduction to Information Retrieval*, Cambridge University Press, 2009, pp. 253–287.

[7] B. Irie and Sei Miyake, "Capabilities of Three-layered Perceptrons," *IEEE nternational Conf. Neural Networks*, pp. 641–648, 1988.

[8] J. Fierrez, A. Morales, R. Vera-Rodriguez, and D. Camacho, "Multiple classifiers in biometrics. part 1: Fundamentals and review," *Inf. Fusion*, vol. 44, no. December 2017, pp. 57–64, 2018.

[9] Y. Miao, H. Jiang, H. Liu, and Y. dong Yao, "An Alzheimers disease related genes identification method based on multiple classifier integration," *Comput. Methods Programs Biomed.*, vol. 150, pp. 107–115, 2017.

[10] C. Catal and M. Nangir, "A sentiment classification model based on multiple classifiers," *Appl. Soft Comput. J.*, vol. 50, pp. 135–141, 2017.

[11] M. Pandey and S. Taruna, "Towards the integration of multiple classifier pertaining to the Student's performance prediction," *Perspect. Sci.*, vol. 8, pp. 364–366, 2016.

[12] U. M. Khaire and R. Dhanalakshmi, "Stability of feature selection algorithm: A review," *J. King Saud Univ. - Comput. Inf. Sci.*, 2019.

[13] R. Zhang, F. Nie, X. Li, and X. Wei, "Feature selection with multi-view data: A survey," *Inf. Fusion*, vol. 50, no. May 2018, pp. 158–167, 2019.

[14] M. Moran and G. Gordon, "Curious Feature Selection," *Inf. Sci. (Ny).*, vol. 485, pp. 42–54, 2019.

[15] I. Sumaiya Thaseen and C. Aswani Kumar, "Intrusion detection model using fusion of chi-square feature selection and multi class SVM," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 29, no. 4, pp. 462–472, 2017.

[16] J. Mielniczuk and P. Teisseyre, "Stopping rules for mutual information-based feature selection," *Neurocomputing*, vol. 358, pp. 255–274, 2019.

[17] B. Richhariya and M. Tanveer, "EEG signal classification using universum support vector machine," *Expert Syst. Appl.*, vol. 106, pp. 169–182, 2018.

[18] G. Singh, B. Kumar, L. Gaur, and A. Tyagi, "Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification," *2019 Int. Conf. Autom. Comput. Technol. Manag.*, pp. 593–596, 2019.

[19] X. Fan, L. Wang, and S. Li, "Predicting chaotic coal prices using a multi-layer perceptron network model," *Resour. Policy*, vol. 50, pp. 86–92, 2016.

[20] Y. Quan, Y. Xu, Y. Sun, and Y. Huang, "Supervised dictionary learning with multiple classifier integration," *Pattern Recognit.*, vol. 55, pp. 247–260, 2016.

[21] B. T. Pham, M. D. Nguyen, K. T. T. Bui, I. Prakash, K. Chapi, and D. T. Bui, "A novel artificial intelligence approach based on Multi-layer Perceptron Neural Network and Biogeography-based Optimization for predicting coefficient of consolidation of soil," *Catena*, vol. 173, no. September 2018, pp. 302–311, 2019.

[22] G. Isabelle, W. Maharani, and I. Asror, "Analysis on Opinion Mining Using Combining Lexicon-Based Method and Multinomial Naïve Bayes," *2018 Int. Conf. Ind. Enterp. Syst. Eng. (ICoIESE 2018)*, vol. 2, no. IcoIESE 2018, pp. 214–219, 2019.

[23] K. Bhattacharjee and M. Pant, "Hybrid Particle Swarm Optimization-Genetic Algorithm trained Multi-Layer Perceptron for Classification of Human Glioma from Molecular Brain Neoplasia Data," *Cogn. Syst. Res.*, vol. 58, pp. 173–194, 2019.

[24] D. Chong, N. Zhu, W. Luo, and X. Pan, "Human thermal risk prediction in indoor hyperthermal environments based on random forest," *Sustain. Cities Soc.*, vol. 49, no. April, p. 101595, 2019.

[25] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[26] Kaggle, "Heart Disease." [Online]. Available: https://www.kaggle.com/ronitf/heart-disease-uci/version/1#_=_. [Accessed: 05-Mar-2020].

[27] OpenML, "Breast Cancer." [Online]. Available: https://www.openml.org/d/13. [Accessed: 05-Mar-2020].

[28] S. Xu, "Bayesian Naïve Bayes classifiers to text classification," *J. Inf. Sci.*, vol. 44, no. 1, pp. 48–59, 2018.

[29] sciki-learn, "Tuning the hyper-parameters of an estimator," 2019. [Online]. Available: https://scikit-learn.org/stable/modules/grid_search.html#grid-search. [Accessed: 05-Mar-2020].

[30] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009.

[31] X. Fan and H. Shin, "Road vanishing point detection using weber adaptive local filter and salient-block-wise weighted soft voting," *IET Comput. Vis.*, vol. 10, no. 6, pp. 503–512, 2016.

[32] L. N. Eeti and K. M. Buddhiraju, "A modified class-specific weighted soft voting for bagging ensemble," *Int. Geosci. Remote Sens. Symp.*, vol. November, pp. 2622–2625, 2016.

[33] D. Ruano-Ordás, I. Yevseyeva, V. B. Fernandes, J. R. Méndez, and M. T. M. Emmerich, "Improving the drug discovery process by using multiple classifier systems," *Expert Syst. Appl.*, vol. 121, pp. 292–303, 2019.

[34] H. F. Nweke, Y. W. Teh, G. Mujtaba, and M. A. Al-garadi, "Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions," *Inf. Fusion*, vol. 46, no. June 2018, pp. 147–170, 2019.

[35] L. Oliveira, U. Nunes, and P. Peixoto, "On Exploration of Classifier Ensemble Synergism in Pedestrian Detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 1, pp. 16–27, 2010.

[36] R. Susmaga, "Confusion Matrix Visualization," in *Intelligent Information Processing and Web Mining*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 107–116.