International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

IJIEMR Transactions, online available on 6th Jan 2023. Link :http://www.ijiemr.org/downloads.php?vol=Volume-12&issue=Issue 01

DOI: 10.48047/IJIEMR/V12/ISSUE 01/28

Title **A Breast Cancer Diagnosis Method Based on VIM Feature Selection and Hierarchical Clustering Random Forest Algorithm PDF**

Volume 12, ISSUE 01, Pages: 299-308

Paper Authors
1. VELDI SATHWIK RAO, 2. GABBETA SAI SAHITH, 3. BANDARU SRIRAKSHA

A Breast Cancer Diagnosis Method Based on VIM Feature Selection and Hierarchical Clustering Random Forest Algorithm PDF

1. VELDI SATHWIK RAO, Department of CSE, Guru Nanak Institute of Technology,Hyderabad,Telangana,India
sathwikrao4545@gmail.com

2. GABBETA SAI SAHITH, Department of ECE, Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering &Technology. saisahithgabbeta@gmail.com

3. BANDARU SRIRAKSHA, Department of IT, Guru Nanak Institute of Technology, bandarusriraksharao@gmail.com

ABSTRACT: The health of women is seriously threatened by the terrible condition known as neoplastic breast cancer. It is thought to be the main factor in the death from female malignant development. Accurate diagnostic confirmation and effective therapy are crucial for lowering the mortality rate from bosom illness. Strong sickness disclosure has been a common usage of ML methods recently, with Irregular Woods (RF) being one of the most often employed. In any event, throughout the preparation cycle, it is possible to produce choice trees with poor grouping execution and high similarity, which might negatively impact the model's overall arrangement execution. A Hierarchical Clustering Random Forest (HCRF) model is produced as a consequence of this work. The proximity of all the trees is investigated utilising a separate tiered bunching approach for grouping evaluation of selected trees. In order to assemble the numerous levels bunching arbitrary backwoods with high accuracy and low similarity, test trees are contrasted to isolated groups. In addition, we use the Variable Importance Measure (VIM) approach to increase the chosen include number for the bosom malignant growth expectation. The Wisconsin Diagnosis Breast Cancer (WDBC) data bank and the Wisconsin Breast Cancer (WBC) data set from the UCI (College of California, Irvine) AI vault are both used in this work. The preciseness, correctness, responsiveness, explicitness, and AUC of the suggested strategy's presentation are all assessed. Trial findings on the WDBC and WBC datasets reveal that the grouping based on the HCRF calculation using VIM as a component determination technique achieves the maximum exactness, with 97.05% and 97.76%, respectively, when compared to Choice Tree, Adaboost, and Irregular Woods. This study's methodology may be utilised to diagnose bosom illness.

Keywords *– Breast cancer, hierarchical clustering, random forest method, and feature selection are all terms used to describe the disease.*

1. INTRODUCTION

One of the most difficult issues affecting women's health is breast cancer, which affects women more than any other group [1, 2]. According to the most recent global disease estimates for 2020, breast cancer has surpassed cell breakdown in the lungs as the most common disease. A precise and prompt conclusion may reduce bosom disease mortality by increasing patients' chances of receiving effective and convenient treatment [3]. The majority of breast cancer conclusions involve imaging and pathology discoveries. A benign symptomatic strategy that has recently received a lot of attention [4]-[6] is imaging determination rather than pathology conclusion. Nevertheless, imaging findings are frequently anticipated following cancer perceivability and may miss early ID. FNA is a minimally intrusive obsessive-compulsive test that, in light of cell morphology [7], has the potential to yield results with high exactness and low false positive rates. First, the cells from the growth in the bosom are removed with a small needle. The phone's thickness, size, uniformity, perfection, and other characteristics are then measured. Finally, new cases are figured out using the data.

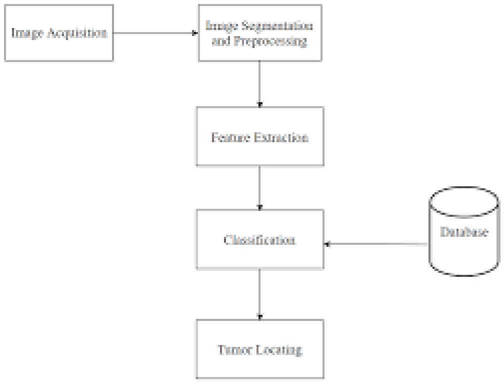


Fig.1: Example figure

A method that uses information to identify inactive data that may not be immediately recognizable is machine learning for FNA information expectation [8]. A well-known group learning technique for overcoming illness is random forest. It differs from other things in two ways: the traits and instances utilized in the decision trees. In terms of the likelihood of overfitting, Random forest performs better than choice trees. In addition, it frequently achieves high expectation precision and is less prone to imbalanced information, noise, and anomalies [9]. Various issues have recently been investigated using irregular woodland [10, 14]. Enhancing feature assurance, changing the majority rule method, setting up the data combination, and fine-tuning the decision tree estimation are all common components of stimulating projects to redesign the erratic forest. On the other hand, fluctuating decision trees in a Random forest classifier simply increase productivity [15].

2. LITERATURE REVIEW

Radiation-induced breast cancer incidence and mortality from digital mammography screening: A modeling study:

Foundation: Mammography screening risk checks for radiation-impelled chest threatening development have not thought about change in segment transparency or illustrative work up following surprising screening revelations. The purpose of this study was to evaluate the distribution of radiation-induced breast disease occurrence and mortality from computerized mammography screening, taking into account openness from screening and indicative mammography as well as variations in measurements between women. The arrangement made use of two different ways to show divertissements. The setting is the US populace. Patients: Women between the ages of 40 and 74. Mediation: Advanced mammography screenings every year or every half year for women 40, 45, or 50 to 74. The estimates include the number of lifetime breast cancer deaths (benefits) and radiation-induced breast cancer deaths (harms) per 100,000 screened women. Results: In contrast to the 968 breast cancer deaths that were prevented by screening, the annual screening of 100,000 women between the ages of 40 and 74 was expected to result in approximately 125 new cases (95 percent confidence interval, 88 to 178) and 16 deaths (CI, 11 to 23). It was anticipated that women in the 95th percentile would develop 246 cases of radiation-induced breast cancer, or 32 deaths per 100,000 women. It was anticipated that women with large breasts, which account for 8% of the population, would have a higher risk of radiation-induced breast cancer than other women (113 disease cases and 15 deaths per 100,000 women). Beginning at age 50, semiannual screening decreased the risk of disease caused by radiation by five times. Radiation-induced bosom disease resulted in a loss of life years that could not be estimated. End: The rate and mortality of radiation-induced breast cancer from advanced mammography screening are all affected by screening portion variation, demonstrative stir up, beginning age, and screening recurrence. Radiation-induced bosom disease may be more difficult for women with large stomachs to treat. The primary funders are the US Preventive Services Team, the Public Disease Foundation, and the Office for Medical Care Exploration and Quality.

Radiomics and machine learning with multiparametric breast MRI for improved diagnostic accuracy in breast cancer diagnosis:

The objective of this multicenter review study was to improve bosom malignant growth analysis by autonomously and as a multiparametric X-ray using machine learning

(ML) of dynamic contrast-enhanced (DCE) and diffusion-weighted imaging (DWI) radiomics models. Patients with a thought further creating chest advancement on chest X-pillar classed as BI-RADS 4 and who later had picture facilitated biopsy were coordinated (Acknowledgment Sloan Kettering Ailment Center, January 2018-Walk 2020; January 2011-August 2014: There were 104 wounds, with a mean size of: Clinical College Vienna 22.8 mm; range: 7 to 99 mm) in 93 people (ages: 49 years and 12 months; 46 (all women) are harmful and 58 (safe). Radiomics credits were figured. After that, a multivariable model incorporating the five most significant characteristics revealed major strength areas for a method for identifying cancers that pose no threat from those that do. For each method, a five-crease cross-approved medium Gaussian support vector machine (SVM) model was developed. The AUC of a model with highlights removed by DWI was 0.79 (95% CI: 0.70-0.88), whereas the AUC of a model with highlights separated by DCE was 0.83 (95% CI: 0.75-0.91). The highest AUC is a multiparametric radiomics model with DCE and DWI inferred qualities was discovered. 95% CI: between 0.77 and 0.92), and the accuracy of the conclusion 95% CI: 73.0-88.6). Last but not least, radiomics examination in conjunction with multiparametric X-ray ML enables a more precise evaluation of thought-upgrading bosom growths that are recommended for biopsy on clinical bosom X-ray, thereby reducing the number of unnecessary benign bosom biopsies and facilitating the accurate discovery of bosom disease.

Machine learning based on multi-parametric MRI to predict risk of breast cancer:

By removing high-throughput characteristics from photos, machine learning (ML) can predict illness. The objective of this work was to develop a nomogram for a multi-parametric MRI (mpMRI) ML model with the intention of anticipating the risk of breast malignant growth. Techniques: The mpMRI was totally associated with T1-weighted imaging (T1WI), T2-weighted imaging (T2WI), clear dissemination coefficient (ADC), Ktrans, Kep, Ve, and Vp. The updated T1WI map's clarified points of interest were planned in accordance with various guides in each cut. On each boundary map, there were 1,132 highlights and the top ten primary parts. Ten rounds of five-crease cross-approval were performed on the ML models, one for each of the single and multiple parameters. The decision to select the model with the highest area under the curve (AUC) was confirmed by the adjustment and choice bends. The best ML model and the patient attributes were utilized to make the nomogram. During the investigation, 144 harmful growths and 66 benign sores were discovered. Individually, patients with benign and dangerous growths had an extremely high average age of 42.5 and 50.8 years. The remaining Ktrans components were given lower priority than the sixth and fourth ones. For non-upgraded T1WI, improved T1WI, T2WI, and ADC models, the Ktrans, Kep, Ve, and Vp AUCs were 0.86, 0.81, 0.81, 0.83, 0.79, 0.81, 0.84, and 0.83, respectively. The alignment and choice bends were used to verify that the best model had an AUC of 0.90. A nomogram for the likelihood of a malignant growth in the bosom was created using the patient's age and the best ML models. Decision: A nomogram could raise expectations for preoperative bosom malignant growth.

Fine-needle aspiration cytology of colloid carcinoma breast in correlation with histopathology:

Fine-needle objective biopsy has for a long while been used to break down and treat unquestionable chest injuries. A fascinating form of breast cancer with distinct cytological and histological characteristics is colloidal carcinoma, also known as pure mucinous carcinoma. Although fine-needle aspiration specimens of mucinous carcinoma of the breast have distinct cytologic characteristics, little research has been done on the relationship between these characteristics and cytologists' ability to clearly identify this growth. Our case study is about a female patient who is 78 years old. Histology confirmed the diagnosis of mucinous carcinoma of the breast made by cytology.

Reviewing ensemble classification methods in breast cancer:

Multiple approaches to dealing with a similar problem are combined in gathering strategies. This strategy was developed to improve the assets of individual techniques while compensating for their shortcomings. In a variety of fields, including bioinformatics, group methods are currently frequently used to carry out forecasting tasks like order and relapse. Breast cancer, which accounts for the majority of female deaths and is the most common type of cancer, has become the focus of research efforts by medical professionals. This review aims to investigate the most cutting-edge methods of group planning in relation to breast cancer in nine areas: distribution scenes, medical projects that were handled, observational and research methods used, types of groups proposed, single procedures used to build the groups, an approval system used to evaluate the outfits proposed, devices used to build the groups, and improvement strategies for the single methods. As part of a study on effective planning, this report was written. Results Four internet-based data sets were used to examine a total of 193 distributions that started around 2000: Scopus, PubMed, IEEE Xplore, and the computerized library of the ACM This study found that out of the six currently available clinical tasks, the demonstrative clinical task was the most frequently investigated, and the trial-based experimental type and assessment-based research type were the most frequently used systems in the selected investigations. For grouping assignments, the homogeneous type was used the most. Choice trees, support vector machines, and fake brain networks were found to be the single systems that were used the most frequently to create gathering classifiers in this planning study. The Wisconsin Bosom Malignancy dataset was the appraisal system that analysts used most frequently to lead their trials, while k-overlay cross-approval was the most important approval method. Weka and R Writing computer programs are two instruments that may be used to coordinate examinations with company request estimations. The framework search method was the most frequently used to change the boundary settings of a single classifier, but few studies considered streamlining the single procedure from which their recommended gathering was constructed. The conclusion of this investigation provides a comprehensive look at how troupe techniques are used to treat breast cancer. We provide bosom disease researchers with ideas as a result of our discoveries, which demonstrate that there are various gaps and concerns. In addition, when we looked at the distributions that were found in our orderly planning investigation, we found that, in comparison to single classifiers, most of them made excellent discoveries regarding gathering classifier execution. A comprehensive writing survey and meta-examination, followed by a top-to-bottom examination to demonstrate the prevalence of troupe classifiers over conventional methodologies, will be required to gather the information presented in the writing.

3. METHODOLOGY

Random Forest (RF) has acquired ubiquity lately as an ML framework able to do really diagnosing a wide scope of sicknesses. Be that as it may, choice trees with unfortunate grouping execution and high similitude might be shaped during the preparation stage, influencing the model's general characterization execution.

Disadvantages:

1. Imaging diagnoses typically need to be confirmed after the tumour has been spotted, and they may miss early detection.
2. Machine learning is a way of analysing data to discover latent knowledge that may not be obvious at first.

In this paper, a Hierarchical Clustering Random Forest (HCRF) model is created. A choice tree grouping investigation is carried out by evaluating how comparable one choice tree is to the others using the progressive bunching method. To form the progressive bunching arbitrary timberland, delegate trees are selected from split groups with great precision. Using the Variable Importance Measure (VIM) approach, we also enhance the selected include number for the breast malignant growth expectation. The Wisconsin Breast Cancer (WBC) and Wisconsin Diagnosis Breast Cancer (WDBC) datasets from the UCI ML vault were used in this review.

Advantages:

1. High accuracy and little similarity.
2. This study's method is a useful tool for detecting breast cancer.

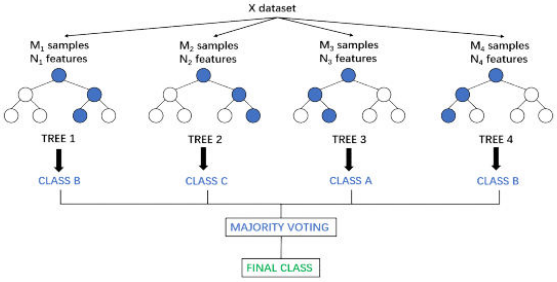


Fig.2: System architecture

MODULES:

To carry out the aforementioned project, we created the modules listed below.

- Information investigation: We will stack information into the framework utilizing this module.
- Handling: We will peruse information for handling utilizing this module.
- Dividing information into train and test: We will partition information into train and test utilizing this module.
- Model generation: Model building - Decision tree, adaboost classifier, random forest, HCRF-using extratree, SGD classifier, and voting classifier.
- User signup and login: Using this module will result in registration and login.
- User input: Using this module will result in prediction input.
- Prediction: the final predicted value will be presented.

4. IMPLEMENTATION

The following algorithms were utilised in this research.

Decision tree:

A non-parametric managed learning strategy known as a decision tree can be used for both characterization and relapse applications. It has a progressive tree structure with a root hub, branches, inner hubs, and leaf hubs.

adaboost classifier:

The AdaBoost calculation, short for Versatile Helping, is a Supporting methodology utilized in ML as a Group Strategy. Versatile Supporting is so named in light of the fact that the loads are reassigned to each occurrence, with bigger loads applied to erroneously classified cases.

Random forest:

An Random Forest Technique is a directed Machine learning (ML) calculation that is generally utilized in ML for characterization and relapse issues. We realize that a backwoods is comprised of many trees, and the more trees there are, the more vivacious the timberland is.

HCRF-using extratree:

Like the random forests procedure, the extra trees calculation creates an enormous number of decision trees, yet the examining for each tree is arbitrary and without substitution. This produces a dataset with one of a kind examples for each tree. For each tree, a specific number of elements are picked indiscriminately from the total assortment of highlights.

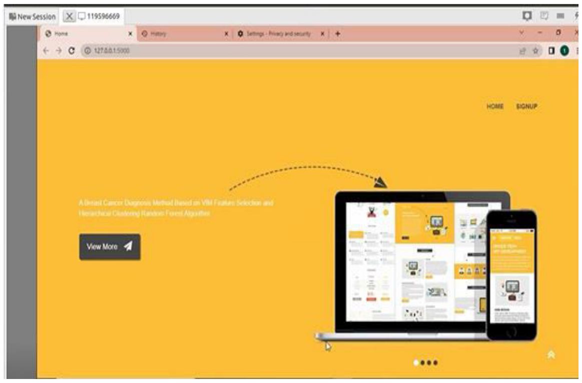
SGD classifier:

The SGD Classifier is a straight classifier (SVM, logistic regression, etc) that has been enhanced utilizing the SGD. These are two particular thoughts. Calculated Relapse or direct Support Vector Machine is an ML calculation/model, while SGD is an enhancement approach.

Voting classifier:

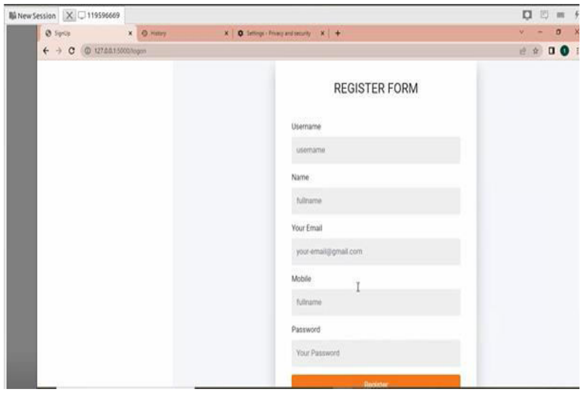
Voting Classifier is an ML approach that Kagglers frequently use to work on the exhibition of their model and ascend the position stepping stool. Voting Classifier may likewise be utilized to increment execution on genuine world datasets, despite the fact that it has huge restrictions.

5. EXPERIMENTAL RESULTS



A Breast Cancer Diagnosis Method Based on VIM Feature Selection and Hierarchical Clustering Random Forest Algorithm

Fig.3: Home screen



REGISTER FORM

Username
username

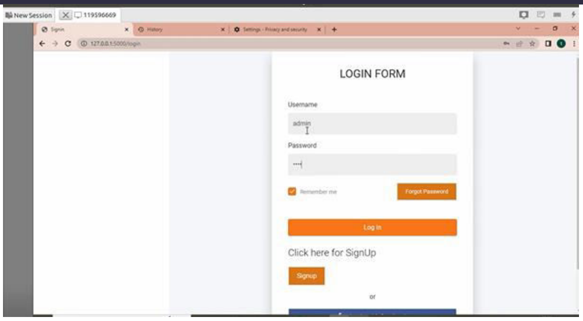
Name
fullname

Your Email
your-email@gmail.com

Mobile
fullname

Password
Your Password

Fig.4: Registration



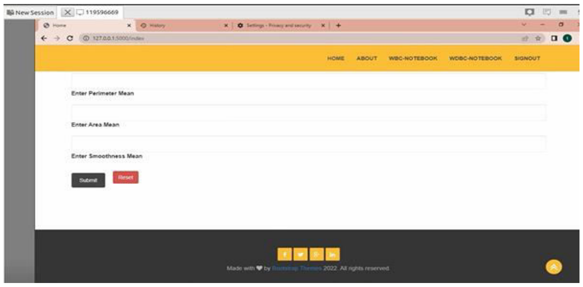
LOGIN FORM

Username: admin

Password: ••••

Click here for SignUp

Fig.5: Login



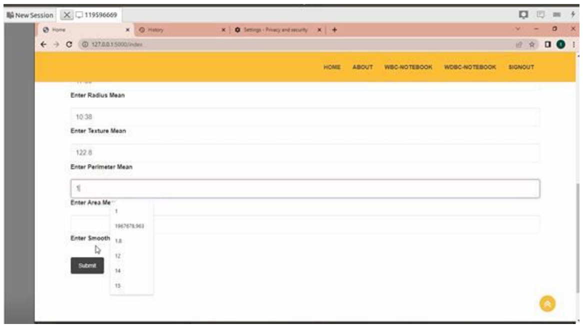
Enter Perimeter Mean

Enter Area Mean

Enter Smoothness Mean

Made with ♥ by Bootstrap Themes 2022. All rights reserved.

Fig.6: Main screen



Enter Radius Mean: 10.38

Enter Texture Mean: 122.8

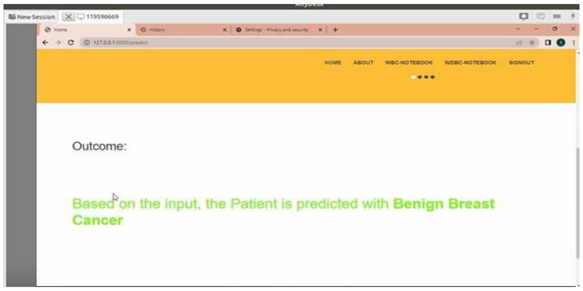
Enter Perimeter Mean: 1

Enter Area Me

Enter Smooth

1
1967678.963
1.8
12
14
15

Fig.7: User input



Outcome:

Based on the input, the Patient is predicted with **Benign Breast Cancer**

Fig.8: Prediction result

6. CONCLUSION

At last, we made a model for bosom disease determination that utilizes VIM for include choice and HCRF for grouping. Both of these procedures not just work on the classifier's exhibition and speculation limit, however they additionally bring down the model's intricacy and testing time. At long last, on the WDBC dataset, our proposed strategy acquires 97.05% exactness and on the WBC dataset, it accomplishes 97.76% precision. When contrasted with the typical irregular woodland model, the proposed HCRF model further improves exactness on the WDBC and WBC datasets by 0.68 percent and 0.5 percent, respectively. In reality exhibition, this is significant on the grounds that it shows that more chest sickness can be distinguished early and more lives can be saved. For creating basic assortment with various types of focal understudies, such as mind associations and support vector machines, as well as other social occasion learning methods, our proposed strategy has a high reference motivation. The proposed method has some clinical implications for the identification of breast cancer because it could also be used to identify various types of malignant growth and provide clinicians with early demonstrative guidance. For individuals who have a history of bosom disease, such a model may lead to the most reasonable treatment and a shorter mediation. In order to work on the variety of random forest decision trees, we will need to demonstrate the decision trees and decompose the underlying assortment in the future. In addition, in order to improve our method's mental prowess, we will employ heuristic strategies to alter significant boundaries.

REFERENCES

[1] R. L. Siegel, K. D. Miller, and A. Jemal, ‘‘Cancer statistics, 2017,’’ CA Cancer J. Clinicians, vol. 60, no. 1, pp. 277–300, 2015.

[2] L. Chang, L. S. Weiner, S. J. Hartman, S. Horvath, D. Jeste, P. S. Mischel, and D. M. Kado, ‘‘Breast cancer treatment and its effects on aging,’’ J. Geriatric Oncol., vol. 10, no. 2, pp. 346–355, Mar. 2019.

[3] H. Danish and S. Goyal, ‘‘Early diagnosis and treatment of cancer series: Breast cancer,’’ Int. J. Radiat. Oncol. Biol. Phys., vol. 80, no. 3, pp. 956–957, 2011.

[4] D. L. Miglioretti, J. Lange, J. J. Van Den Broek, C. I. Lee, and R. A. Hubbard, ‘‘Radiation-induced breast cancer incidence and mortality from digital mammography screening: A modeling study,’’ Ann. Internal Med., vol. 164, no. 4, pp. 205–214, Jan. 2016.

[5] I. D. Naranjo, P. Gibbs, J. S. Reiner, R. Lo Gullo, C. Sooknanan, S. B. Thakur, M. S. Jochelson, V. Sevilimedu, E. A. Morris, P. A. T. Baltzer, T. H. Helbich, and K. Pinker, ‘‘Radiomics and machine learning with multiparametric breast MRI for improved diagnostic accuracy in breast cancer diagnosis,’’ Diagnostics, vol. 11, no. 6, p. 919, May 2021.

[6] W. Tao, M. Lu, X. Zhou, S. Montemezzi, G. Bai, Y. Yue, X. Li, L. Zhao, C. Zhou, and G. Lu, ‘‘Machine learning based on multi-parametric MRI to predict risk of breast cancer,’’ Frontiers Oncol., vol. 11, p. 226, Feb. 2021.

[7] D. Maruti and G. Vandana, ‘‘Fine-needle aspiration cytology of colloid carcinoma breast in correlation with histopathology,’’ Apollo Med., vol. 12, no. 4, pp. 264–266, Dec. 2015.

[8] David and Edwards, ‘‘Data mining: Concepts, models, methods, and algorithms,’’ J. Proteome Res., vol. 2, no. 3, p. 334, 2003.

[9] M. Hosni, I. Abnane, A. Idri, J. M. C. de Gea, and J. L. F. Alemán, ‘‘Reviewing ensemble classification methods in breast cancer,’’ Comput. Methods Programs Biomed., vol. 177, pp. 89–112, Aug. 2019.

[10] J. Gómez-Ramírez, M. Ávila-Villanueva, and M. Á. Fernández-Blázquez, ‘‘Selecting the most important self-assessed features for predicting conversion to mild cognitive impairment with random forest and permutationbased methods,’’ Sci. Rep., vol. 10, no. 1, pp. 1–15, Dec. 2020.