Contributing to diagnoses of Mental Disease Using New Optimization Machine Learning Methods

Mustafa Adil Fayez
Department of ECE, Institute of Science, Altınbaş University, Istanbul Turkey
Email: mustafa.fayez@ogr.altinbas.edu.tr

Abstract: Mental illness is known to be difficult to diagnose and we can say that if someone has a mental illness, it can affect someone for years before diagnosis. Geriatricians often encounter a large number of patients for treatment without being able to reduce or automatically diagnose them. Designing a system to help better to diagnose this disease and to reduce cost and time is our aim in this study. We used a mental dataset applied with data mining optimization algorithms and we applied it with the Python programing language, including training the test split and pre-processing feature selection model used random forest (RFE) and feature importance to enhance the system results and accuracy for mental dataset in our research and processed the missing values that found with attributes.The best accuracy was achieved by Adaboost optimization model, which gave us 99% accuracy and the Adaboost ensemble merged with the decision tree produced a 94% accuracy. Moreover, the random forest optimization produced better accuracy at 96% and 92% resulted from using the SVM algorithm. Finally, this optimization system by merging the two algorithms to work together will be more efficient and better able to help classify and diagnose suffering patients using a huge amount of data in little time and at low cost.

Keywords: Mental disease, Ensemble techniques, Python, Data mining

1. Introduction

A mental illness is a condition that is considered to be a clinically important illness for the individual [1], the instruction of emotions, or it is behaviour that reflects the imponderables in the psychological, biological or developing processes that make his mental function. Mental illnesses are generally accompanied by considerable suffering in social, working, or other important situations [3]. A predictable or culturally appropriate reaction to popular stress or injury, such as a loved one's loss or serious disease, is not a mental illness. Informally different behaviour (e.g., political, religious or sexual) and disputes that are essentially between the separate and the social order are not mental illnesses unless the deviation or inconsistency of an individual's dysfunction [4].Alzheimer's Disease is a multifaceted disease that results in cumulative cerebrovascular disease leading gradually to cognitive deterioration and eventually dementia [5]. Big data sciences for research have become increasingly important for health care, but mental health applications have so far been relatively limited. The challenges of big data traditionally represent volume challenges (dataset size), speed (real-time data acquisition), diversity (multiple sources and types) and newly added factors of "variance" and "credibility." Reproduce the undependability of emerging information from some sources [6].

Mental disease is can be defined by a great degree of weakness, such as emotional distress, which leads to unhappiness and various neurological illnesses of which 25% are experienced in industrialized and developing countries [7]. Terabytes of data exist, 80% of which are unorganized, so it is hard to address with database management tools or other modern technologies. Approximately $2.3 trillion is spent internationally on the treatment of mental disease [8]. By refining the superiority of handlings, we can of course decrease costs and improve this superiority by introducing tools and techniques of data extraction from mental conditions [9]. There is a multiplicity of mental illnesses (including dementia, depression, bipolar disorder and Alzheimer's disease) that are dementia-derived diseases [10]. In this study, we put forward the research question as to whether there are efforts linked to data mining methods being applied to mental disease with the aim of diagnosing mental pathologies. Therefore, the goal of our work is to provide an optimization machine learning techniques and communicable diseases for mental illness by merging some machine learning algorithms together in order to enhance the prediction system results and accuracy with less time. This comprehensive study is the main contribution of our work and it allows us to conduct continuous future studies in the establishment of ways to diagnose mental disease.

2. Related Works

Ami Robert Stewart and Katharina Davis [6] in 2016 were working on big data in the current state of mental illness research and emerging possibilities. They had sought to review and use large data resources in this disorder by identification of the features of technologies to date and consideration of potential innovation patterns. Clear differences were evident in the covered geographical areas and in disturbances and interventions of great interest.

Jiang Bian, Laura E. Barnes and Guanling Chen [11] in 2017 conducted research into the discovery of illnesses consuming electric health data and linear differential analysis. They made a broad assessment using the CRDA assessment and a large-scale, real-world EHR data collection to forecast mental conditions (such as depression and anxiety) among university students from ten US universities.

Donnelly-Kiho, Patricio Andresi, Eva Pascarello, Guido Orlando and Gomez [12] in 2017 conducted research on morphometric signatures for Alzheimer's Disease using automated learning techniques. They proposed separating participants into three clusters according to the State Mental Examination Standard (MMSE).

Stephan Feder and Benedikt Sundermann [13] in 2017 worked on the homogeneity of a sample in unipolar depression. The subjects were evaluated using functional communication analyzes dominated by the general effects of the disease (1), and (3). Exploratory cluster analyses exposed two weakly subdivided subcategories of unhappy patients. These subclasses differed in terms of mean duration of despair and the percentage of patients with simple depressive symptoms and anxiety at one time.

Mahdi Ben Youssef, Abdul Latif Al-Abyad and Hind Al-Hadiri [14] in 2017 were working on the approach of diagnosing Alzheimer's disease using data mining techniques. They proposed a model consisting of three classification and prediction algorithms: decision trees, a discriminatory analysis and a logistic regression.

Susil G. Alonso, Isabel de la Torre-Díez, Lola M. Nozaleda and Manuel Franco [15] in 2018 suggested DM methods and techniques in mental illness as a methodical survey. They found that 211 materials linked to machine learning algorithm functional to major mental diseases were found. 72 articles were known as related work, 32% were Alzheimer's, 22% were dementia, 24% were depression, 14% were schizophrenia and 8% were polar disorders.

Robert L. Spitzer, et al [16] in 2018 conducted their work on medical and mental disorders. They proposed a set of criteria to identify medical illness that receipts into account the difficulties of the problematic and is abnormally similar in its classification to circumstances traditionally considered medical complaints, including mental illnesses.

In these previous researches, we found that they used normal systems to deal with mental disorder detection problem. While designing high efficiency optimization machine learning system using Python programming language including better pre-processing methods as feature selection model used random forest (RFE) and feature importance in enhancing system outcomes and accuracy for mental data set in our research and processing the missing values found with attributes.

3. Methodology

In this work, we deal with data mining classification techniques to obtain a powerful system for the purpose of diagnosing mental disease with the application of a mental dataset.

3.1. Mental Disease Dataset

This 2014 study dataset tracks attitudes to mental disease and the recurrence of mental illness at work. We obtained this dataset from the kaggle datasets website [17]. The result attribute is (work-interfere) which states: Would he find that it interferes with work if there is mental illness? "NA,""Never,""Rarely,""Often" and "Often are the potential answers to this issue it can be deduced from this query if anyone has a mental disorder, even though the question is not expressly asked in the study. They may have a mental condition if anyone picks something other than "NA" They may or may not have one if they choose "NA".

3.2. Data Pre-processing and Future Selection

Clinical datasets often have the problem of missing values, and this can affect the accuracy of the system [18]. We have processed the problem of missing values according to the scientific solution by replacing them with the mean values of each attribute [19]. We translated the categorical attributes into numerical values by using a label encoder suitable for work with Python and we applied the scalar normalization method [20] to mental dataset. Moreover, we used the feature importance with the random forest model selection [21]. The accuracy of the feature selection (model selection) was 94% and we eliminated five attributes from the mental dataset (timestamp, country, state, no_employees and comments). Moreover, we removed these attributes from the dataset to enhance the results and accuracy of the mental diagnoses system.

.3. Data Mining Tools

– Random Forest Optimization

RF is essentially a group of decision trees, each of which classifies the dataset utilizing a subset of variables (often linearly)[22]. The number of forest trees and the number of subgroup variables are super parameters and must be chosen a priori [23]. Compared to the average number of variables, the number of trees is in the hundreds, whereas the collection of variables is very tiny. Also have a standard way of determining the validity of input variables (forecasters)[24]. We used optimization of Random forests and the better parameters were {'bootstrap': False, 'max_depth': None, 'min_samples_split': 2, 'n_estimators': 393}.

– Support Vector Machine (SVM)

The SVM is an automated learning algorithm that analyzes estimation data and performs regression analysis [25]. The SVM is a controlled learning method that searches for and arranges sorted data with margins between the two as far apart as possible [26]. SVM files are used to classify text, classify images and recognize handwriting. There are different kinds of kernels, such as linear, RBF, sigmoid and poly models. We used the RBF kernel with the mental dataset as it provided the best results with high accuracy with data mining techniques used in mental diagnosis systems [27].

– Ada Boost Optimization

Ada Boost combines a weak classification technique to shape a powerful workbook. One technique may distinguish a target poorly [28]. However, if we combine multiple actions with the selection of the training set for each frequency and assign the appropriate weight to the final vote, we can obtain a good degree of accuracy for the overall rating [29]. We used Adaboost optimization and we got best parameters like {'learning_rate': 0.001, 'n_estimators': 100}. We used Ada Boost ensemble [29] with the Decision Tree to improve and enhance the results of diagnosis systems and achieve high accuracy.

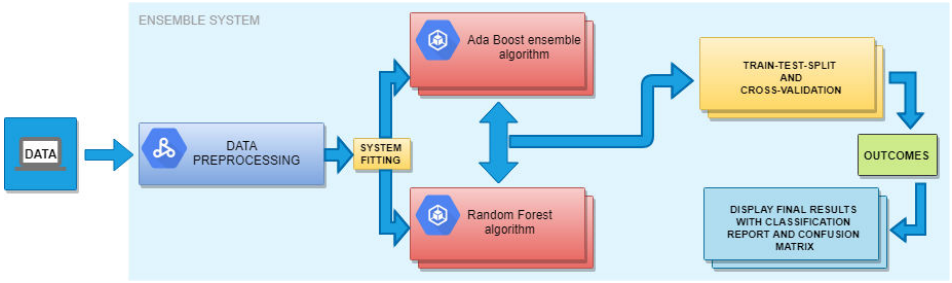

Fig. 1. Ensemble system structure.

4. Results and Discussion

In our mental diagnosis systems, we used different kinds of data mining techniques applied with the Python programing language. Moreover, we used the training test split [30] to divide our dataset into training and testing with (5-15) Random State and we used 30% of the mental dataset as a testing size and 70% as the training size. Additionally, we used pre-processing for our dataset, including feature selection and missing values pre-processing, which changed and enhanced our accuracy. Different results and accuracy were obtained with 22 attributes of mental disease, such as 96% using the Random Forest optimization and 99% using the Adaboost optimization. In addition, we obtained 92% accuracy using the SVM classifier with the rbf kernel and 94% using the Adaboost ensemble with the decision tree and extra tree algorithms. This study has been accurate with the mental dataset shown in Table 2 below, which shows the different accuracies of mental disease diagnosis systems.

Table 2. Accuracy of the dataset in our study

	Random Forest Optimization	Adaboost Optimization	SVM	Adaboost Ensemble with the Decision Tree
Mental Dataset	96%	99%	92%	94%

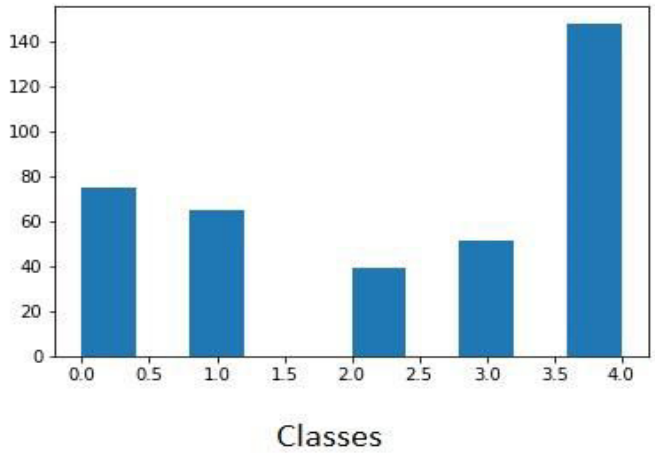

Fig. 2. Distribution of outcome classes

In the figure above, we see the distribution of the mental prediction attribute (work_interfere). We acquired the level of class number four as being the highest of all the classes, which means the answer is "sometimes." The tables below present classification reports of our system results with data mining techniques.

Table 3. Results of the RF optimization with the mental dataset.

Class No.	Precision	Recall	F1. score	Support
NA (Class 0)	0.99	1.00	0.99	74
Never (Class 1)	0.90	0.94	0.92	68
Often (Class 2)	0.86	0.88	0.87	41
Rarely (Class 3)	1.00	0.91	0.95	56
Sometimes (Class 4)	1.00	1.00	1.00	139
Avg. / Total	0.96	0.96	0.96	378

Table 4. Results of the Adaboost optimization.

Class No.	Precision	Recall	f1. score	Support
NA (Class 0)	1.00	0.99	0.99	73
Never (Class 1)	0.97	1.00	0.98	59
Often (Class 2)	1.00	0.98	0.99	42
Rarely (Class 3)	1.00	1.00	1.00	56
Sometimes (Class 4)	1.00	1.00	1.00	148
Avg. / Total	0.99	0.99	0.99	378

We found that the results and accuracy using the Adaboost Ensemble merged with the Decision Tree algorithms were better than using each technique in a separate form.

Table 5. Results of the Adaboost Ensemble with the Decision Tree.

Class No.	Precision	Recall	F1. score	Support
NA (Class 0)	0.97	1.00	0.99	73
Never (Class 1)	0.92	0.97	0.94	59
Often (Class 2)	0.89	0.79	0.84	42
Rarely (Class 3)	0.87	0.80	0.83	56
Sometimes (Class 4)	0.97	1.00	0.99	148
Avg. / Total	0.94	0.94	0.94	378

Table 6. Results of the Support Vector Machine (SVM).

Class No.	Precision	Recall	F1. score	Support
NA (Class 0)	0.93	0.96	0.95	73
Never (Class 1)	0.90	0.90	0.90	59
Often (Class 2)	0.95	0.88	0.91	42
Rarely (Class 3)	0.93	0.75	0.83	56
Sometimes (Class 4)	0.92	0.99	0.95	148
Avg. / Total	0.92	0.92	0.92	378

5. Conclusion

This tudy pertains to determining the best system for the diagnosis of mental disease. We used the Python programing language to apply data mining classification techniques to classify the mental dataset with 22 attributes. We achieved better accuracy using the Adaboost optimization technique at 99%, and with the Adaboost Ensemble with the Decision Tree, the accuracy was 94%. With the Random Forest optimization, we had 96% and 92% when using the support vector machine (SVM) technique. Moreover, we achieved the highest accuracy as we used the training test split, feature importance and selection model such as the Random Forest (RF) pre-processing for the mental dataset in a different manner to produce a better system of diagnosis in minimum time and at low cost.

References

A. J. Wyner, M. Olson, J. Bleich, and D. Mease, "Explaining the success of adaboost and random forests as interpolating classifiers,"*The Journal of Machine Learning Research,* vol. 18, pp. 1558-1590, 2017.

B. T. Pham, D. T. Bui, and I. Prakash, "Landslide susceptibility assessment using bagging ensemble based alternating decision trees, logistic regression and J48 decision trees methods: a comparative study,"*Geotechnical and Geological Engineering,* vol. 35, pp. 2597-2611, 2017.

C. Spiranovic, A. Matthews, J. Scanlan, and K. C. Kirkby, "Increasing knowledge of mental illness through secondary research of electronic health records: opportunities and challenges,"*Advances in Mental Health,* vol. 14, pp. 14-25, 2016.

E. Chiauzzi and A. N. AB, "Patient Perspective."

E. Feczko, N. Balba, O. Miranda-Dominguez, M. Cordova, S. Karalunas, L. Irwin*, et al.*, "subtyping cognitive profiles in Autism Spectrum Disorder using a Functional Random Forest algorithm,"*Neuroimage,* vol. 172, pp. 674-688, 2018.

E. M. Benyoussef, A. Elbyed, and H. El Hadiri, "Data Mining Approaches for Alzheimer's Disease Diagnosis," in *International Symposium on Ubiquitous Networking*, 2017, pp. 619-631.

G. Varoquaux, "Cross-validation failure: small sample sizes lead to large error bars,"*Neuroimage,* vol. 180, pp. 68-77, 2018.

H. Goldstein, J. Carpenter, and M. G. Kenward, "Bayesian models for weighted data with missing values: a bootstrap approach,"*Journal of the Royal Statistical Society: Series C (Applied Statistics),* 2018.

J. Bian, L. E. Barnes, G. Chen, and H. Xiong, "Early detection of diseases using electronic health records data and covariance-regularized linear discriminant analysis," in *Biomedical & Health Informatics (BHI), 2017 IEEE EMBS International Conference on*, 2017, pp. 457-460.

J. E. Patterson and T. M. Edwards, "An introduction to global mental health,"*Families, Systems, & Health,* vol. 36, p. 137, 2018.

K. Polat and U. Sentürk, "A Novel ML Approach to Prediction of Breast Cancer: Combining of mad normalization, KMC based feature weighting and AdaBoostM1 classifier," in *2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, 2018, pp. 1-4. K. Kaplan-Solms, Clinical studies in neuro-psychoanalysis: Introduction to a depth neuropsychology: Routledge, 2018.

L. San, G. Estrada, N. Oudovenko, and E. Vieta, "Rationale and design of the PLACID study: a randomised trial comparing the efficacy and safety of inhaled loxapine versus IM aripiprazole in acutely agitated patients with schizophrenia or bipolar disorder,"*BMC psychiatry,* vol. 17, p. 126, 2017.

L. Squarcina, U. Castellani, M. Bellani, C. Perlini, A. Lasalvia, N. Dusi*, et al.*, "Classification of first-episode psychosis in a large cohort of patients using support vector machine and multiple kernel learning techniques,"*NeuroImage,* vol. 145, pp. 238-245, 2017.

M. Mitchell and M. Billings, "Evaluation of Pharmacist Interventions in Patients with Substance Use Disorder and Mental Illness Managed Through a Collaborative Telehealth Educational Model," 2018.

M. P. Dooshima, E. N. Chidozie, B. J. Ademola, O. O. Sekoni, and I. P. Adebayo, "A Predictive Model for the Risk of Mental Illness in Nigeria Using Data Mining,"*International Journal of Immunology,* vol. 6, p. 5, 2018.

M. Mursalin, Y. Zhang, Y. Chen, and N. V. Chawla, "Automated epileptic seizure detection using improved correlation-based feature selection with random forest classifier,"*Neurocomputing,* vol. 241, pp. 204-214, 2017.

M. Feres, Y. Louzoun, S. Haber, M. Faveri, L. C. Figueiredo, and L. Levin, "Support vector machine-based differentiation between aggressive and chronic periodontitis using microbial profiles,"*International dental journal,* vol. 68, pp. 39-46, 2018. P. Dhaka and R. Johari, "Big data application: Study and archival of mental health data, using MongoDB," in *Electrical, Electronics, and Optimization Techniques (ICEEOT), International Conference on*, 2016, pp. 3228-3232.

P. A. Donnelly-Kehoe, G. O. Pascariello, J. C. Gómez, and A. D. N. Initiative, "Looking for Alzheimer's Disease morphometric signatures using machine learning techniques,"*Journal of neuroscience methods,* vol. 302, pp. 24-34, 2018.

R. L. Spitzer, J. Endicott, and J.-A. M. Franchi, "Medical and mental disorder: Proposed definition and criteria," in *Annales Médico-psychologiques, revue psychiatrique*, 2018, pp. 656-665.

R. Stewart and K. Davis, "'Big data'in mental health research: current status and emerging possibilities,"*Social psychiatry and psychiatric epidemiology,* vol. 51, pp. 1055-1072, 2016.

S. K. A. LEE, "Classification of SmartMentalTech Services and Application for Comprehensive Mental Healthcare Stepped-Care Model (CMHSCM): Health Psychological Approach,"*Procedia Computer Science,* vol. 141, pp. 302-310, 2018.

S. Feder, B. Sundermann, H. Wersching, A. Teuber, H. Kugel, H. Teismann*, et al.*, "Sample heterogeneity in unipolar depression as assessed by functional connectivity analyses is dominated by general disease effects,"*Journal of affective disorders,* vol. 222, pp. 79-87, 2017.

S. G. Alonso, I. De La Torre-Díez, S. Hamrioui, M. López-Coronado, D. C. Barreno, L. M. Nozaleda*, et al.*, "Data mining algorithms and techniques in mental health: a systematic review,"*Journal of medical systems,* vol. 42, p. 161, 2018.

S. Ramírez-Gallego, B. Krawczyk, S. García, M. Woźniak, and F. Herrera, "A survey on data preprocessing for data stream mining: Current status and future directions,"*Neurocomputing,* vol. 239, pp. 39-57, 2017.

S. Dimitriadis, D. Liparas, M. N. Tsolaki, and A. s. D. N. Initiative, "Random forest feature selection, fusion and ensemble strategy: Combining multiple morphological MRI measures to discriminate among healhy elderly, MCI, cMCI and alzheimer's disease patients: From the alzheimer's disease neuroimaging initiative (ADNI) database,"*Journal of neuroscience methods,* vol. 302, pp. 14-23, 2018.

T. Shaikhina, D. Lowe, S. Daga, D. Briggs, R. Higgins, and N. Khovanova, "Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation,"*Biomedical Signal Processing and Control,* 2017.

Y.-K. Hou, H. Chen, C.-Y. Xu, J. Chen, and S.-L. Guo, "Coupling a Markov Chain and Support Vector Machine for at-site downscaling of daily precipitation,"*Journal of Hydrometeorology,* vol. 18, pp. 2385-2406, 2017.

(2014). *Survey on Mental Health in the Tech Workplace in 2014*. Available: https://www.kaggle.com/osmi/mental-health-in-tech-survey/data