Indonesian Journal of Electrical Engineering and Informatics (IJEEI)
Vol. 9, No. 4, December 2021, pp. 929~942
ISSN: 2089-3272, DOI: 10.52549/ijeei.v9i4.3218

Utilising Target Adjacency Information for Multi-target Prediction

Ruhaila Maskat[1], Ramli Musa[2], Norizah Ardi[3], Noor Afni Deraman[4], Zaaba Ahmad[5], Wang Qingchen[6], Shukor Sanim Mohd Fauzi[7], Ray Adderley JM Gining[8], Tajul Rosli Razak[9]
[1,5]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Selangor, Malaysia
[2]Department of Psychiatry, Kulliyyah of Medicine, International Islamic University Malaysia
[3]Academy of Language Studies, Universiti Teknologi MARA, Selangor, Malaysia
[4]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Melaka, Malaysia
[6]Faculty of Business and Economics, University of Hong Kong, Hong Kong
[7,8,9]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Perlis, Malaysia

Article Info

Article history:

Received Jul 13, 2021
Revised Sep 6, 2021
Accepted Nov 13, 2021

Keyword:

Multi-target prediction
Side information
Machine learning
Mental illnes

ABSTRACT

In this paper, we explored how information on the cost of misprediction can be used to train supervised learners for multi-target prediction (MTP). In particular, our work uses depression, anxiety and stress severity level prediction as the case study. MTP describes proposals which results require the concurrent prediction of multiple targets. There is an increasing number of practical applications that involve MTP. They include global weather forecasting, social network users' interaction and the thriving of different species in a single habitat. Recent work in MTP suggests the utilization of "side information" to improve prediction performance. Side information has been used in other areas, such as recommender systems, information retrieval and computer vision. Existing side information includes matrices, rules, feature representations, etc. In this work, we review very recent work on MTP with side information and propose the use of knowledge on the cost of incorrect prediction as side information. We apply this notion in predicting depression, anxiety and stress of 270,322 anonymous respondents to the DASS-21 psychometric scale in Malaysia. Predicting depression, anxiety and stress based on the DASS-21 fit an MTP problem. Often, a patient experiences anxiety as well as depression at the same time. This is not unusual since it has been discovered that both tend to co-exist at different degrees depending on a patient's experience. By using existing machine learning algorithms to predict the severity levels of each category (i.e., depression, anxiety and stress), the result shows improved precision with the use of cost matrix as side information in MTP.

Corresponding Author:

Ruhaila Maskat,
Faculty of Computer and Mathematical Sciences,
Universiti Teknologi MARA,
Selangor, Malaysia.
Email: ruhaila@tmsk.uitm.edu.my

1. INTRODUCTION

In traditional prediction, only a single target is used, typically known as single-target prediction (STP). This target can be classified into one of two class labels (binary) or multiple labels (multi-label). Usually, for textual targets classification methods are employed, while for numeric target regression methods are used. The resulting value of a target directly depends on the combination of multiple independent variables that an instance has. The nature of STP can be seen as straightforward and simple. As an example, classifying a patient to either be depressed or otherwise. Conversely, in multi-target prediction (MTP), there is over one target that needs to be predicted at once [1]. Each target can be of differing types (e.g., binary, ordinal, nominal) and can

each either be binary or multi-label [1]. To understand how MTP differs from STP is by extending from the STP example. Here, the aim would be to predict a patient to be jointly experiencing depression and anxiety at differing levels of severity. One scenario is a patient may have mild depression yet extremely severe anxiety. Another scenario is a patient may suffer from depression, but not anxiety. Nevertheless, both the former and latter scenarios require a prediction to be conducted concurrently to determine the level of severity or the absence of it for both targets. Waegeman et al. [1] characterize the basic structure of MTP as been built upon instances X and targets T. A dataset for MTP used for training a model would have an additional component besides X and T to carry the notion of their associations in the form of a set of scores (Y) since X is described by a set of independent variables and T is the set of dependent variables. MTP aims to predict Y for every instance-target (X, T) pair. Y can be of nominal, ordinal or real values. Therefore, for every n number of instances and m number of targets, the size of Y would be the matrix of $n \times m$.

An increasing realisation of MTP's usage in a wide range of domains is shifting MTP to be at the centre stage for current prediction tasks [2]. As interactions between entities in the real world become increasingly complex, prediction tasks must be adept to handle such complexity without compromising performance. Recent applications of MTP published include prediction of protein functions in bioinformatics [3], prediction of arch dam deformation in mathematical modelling [4], prediction of soil properties in agriculture [5], prediction of cognitive decline in Alzheimer patients [6] prediction of wheat flour quality parameters [7], prediction of identifying learning styles [8], prediction of drug toxicity [9], prediction of cervical cancer [10] and prediction of wine category [11]. These studies discovered that using MTP instead of STP improves performance. Furthermore, MTP reduces overfitting [2].

Under the MTP umbrella, there are a myriad of methods fitting the characteristics of MTP but has long been identified with specific names [1]. They include multivariate regression, multi-label classification, multi-task learning, zero-shot learning and matrix completion. While many existing MTP works use machine learning algorithms designed for STP, this boundary is being pushed with the proposal of new variants of algorithms designed specifically for MTP. They include [2][12-19]. In general, these algorithms were designed to take into account all input features related to all the given targets.

A wide range of MTP techniques has been proposed during the past 10 years. At present, two taxonomies have been built [13]. The first is defined as problem transformation and algorithm adaptation. According to [13], problem transformation turns a multi-target problem into several individual target problems. It has been reported that this approach requires more resources to singly solve each problem and then combine them. On the other hand, algorithm adaptation is deemed more effective in performance as the algorithm is adapted to predict all targets together [14]. With this approach, it has the flexibility to accommodate side information [14]. A more recent taxonomy, which replaced the previous one, classifies MTP as falling into one of these categories: local models and global models [13]. However, the definition and scope remain the same [13].

This study started with a systematic review of the literature. Progress in MTP shows multiple adoptions of side information/knowledge to improve the resulting performance. A variety of side information has since been proposed in different domains and is described in the next subsection.

1.1. Review of State-of-the-art Side Information

The performance of a prediction is of considerable importance. In MTP, it is more complex than STP as all targets must be predicted simultaneously. Nevertheless, the existence of additional information, as described in Waegeman et al. [1] namely side information, can assist in improving MTP's performance. The existence and form of side information depend entirely on the domain problem. In other words, side information may naturally not exist in a domain and if it does exist may take various forms. Side information has been used in many areas: recommender systems [22-23], information retrieval [22-23], computer vision [16] and text summarization [17], to name a few. The common general idea motivating its use is the enhancement that it can bring. There could never be too much side information as potentially many more are yet to be discovered in a variety of domains, each of which provides some unique salient information. To conduct this review, we adopted the Systematic Mapping Study (SMS) approach [18]. The primary steps in SMS are defining research questions, searching, screening found papers, keywording from abstracts and extracting and mapping of data. In this study, we have combined the last two steps as both steps are tightly interlinked, iterating back and forth between each other.

Step 1: Defining of research questions

In this step, we aim to understand the trends of recent works in MTP's side information with relation to the form that it takes, the prediction models that it is devised for and the type of strategy that it is used in.

From this review, researchers and practitioners can gain insights as to the future direction of side information in MTP. To relay our aim, we constructed the following research questions.

RQ1: What are the most recent forms of side information in MTP?
RQ2: What prediction model the side information was devised for?
RQ3: What MTP strategy that the side information was used in?
RQ4: What metrics were used to evaluate the MTP proposals?

Step 2: Conducting a search

In this study, academic publications were collected from different online databases via Google Scholar. Google Scholar's results typically include nearly all online databases i.e. IEEE Xplore, ACM Digital Library and Elsevier ScienceDirect, thus acting as a one-stop-centre. We used the term "multi-target prediction" without adding the terms "side information" or "side knowledge" since their use returned very few results. This could be because both terms were popularised in the field of MTP by Waegeman et al. only in 2019 and have thus not yet gained widespread adoption. Our search involves publications as early as 2020 until the point of writing in 2021. The justification is to provide interested researchers with a quick understanding of where the focus in very recent work lies.

Step 3: Screening found papers

From the 76 results listed in Google Scholar, we examined each paper's title, abstract, keywords, introduction and conclusion to determine their eligibility. Typically, examination of the introduction and conclusion is performed in step 4: Keywording from Abstracts; however, we began this process early to make the subsequent steps easier. Additionally, we devised two exclusion criteria (E) and one inclusion criteria (I):
E1: Duplicates.
E2: Publications that do not clearly use side information.
I1: Papers that use the terms multi-target, multi-variate regression, multi-task learning, zero-shot learning or matrix completion.

Steps 4 & 5: Keywording from abstracts & Extracting and mapping of data

We listed the relevant publications in Table 1. From these publications, we formed unique classification schemes to serve each research question. To this end, data is extracted from the publications and mapped into these schemes. The list of publishers indicates the application of MTP are in various domains (chemistry, medical, bioinformatics, biology, healthcare), not centred solely on computing. This is in alignment with earlier reports on the increasing realisation of MTP's application in multiple domains. It is noteworthy that our list of publications excludes works on MTP that are without the utilisation of side information.

Table 1. List of relevant publications

Publications	Year	Publishers
Breskvar and Džeroski [19]	2021	IEEE Access
Chen et al. [4]	2021	Applied Mathematical Modelling
Santana et al. [5]	2021	Chemometrics and Intelligent Laboratory Systems
Mastelini et al. [2]	2020	Applied Soft Computing
Pliakos and Vens [20]	2020	BMC Bioinformatics
Wu and Lian [21]	2020	Proceedings of the International Joint Conference on Neural Networks
Adıyeke and Baydoğan [13]	2020	Pattern Recognition
Bessadok et al. [22]	2020	Lecture Notes in Computer Science
Liu et al. [23]	2020	IEEE Access
Mignone et al. [24]	2020	Nature Scientific Reports
Liu et al. [25]	2020	Machine Learning for Pharma and Healthcare Applications

RQ1: What are the latest forms of side information in MTP?

The significance of this research question is to learn what forms that side information takes in cutting-edge works. Our study (refer to Table 2) discovered that the majority of recent side information takes the form of matrices to represent the dependencies between targets. Vectors are found next, followed by sequences and rules. A matrix is an intuitive artefact to represent side information as it can capture the value of target pairs within its intersections. The number of dimensions is unlimited thus can easily represent an unlimited number of target pairs. Each matrix typically carries one interdependency information. Cases with more than one interdependency relationship between targets do require more than one matrix. Vectors and sequences are simply subsets of matrices, hence can also support interdependency information well. While these are strictly structured artefacts, decision trees were devised to handle interdependency information with less structure. Rules are obtained from decision trees in light of targets as the final nodes. Interdependency information is domain-specific. Some are a direct representation of principles, methodologies and structures found in the

domain, e.g. drug-target interactions, while others are more implicitly inferred e.g. learning models aimed to predict specific targets.

Table 2. Forms of side information

Publications	Forms
Adıyeke and Baydoğan [13]	A sequence of selected good quality targets.
Bessadok et al. [22]	Matrix of similarity between targets.
Breskvar and Džeroski [19]	Weighted rules generated from an ensemble of decision trees containing target attributes.
Chen et al. [4]	Matrix of highly correlated subsets of targets.
Liu et al. [23]	Vector containing the interdependencies information of targets.
Liu et al. [25]	Matrix of similarity between targets.
Mastelini et al. [2]	Matrix of highly influential targets.
Mignone et al. [24]	Pair of vectors representing the confidence on the existence of interaction between targets.
Pliakos and Vens [20]	Matrix of interactions between targets.
Santana et al. [5]	Matrix of prediction base learner models for each target attribute.
Wu and Lian [21]	Matrix of strongly correlated targets.

RQ2: What prediction model the side information was devised for?

The choice of prediction models is substantially influenced by the trade-off between interpretability and accuracy. A highly interpretable model can provide the reasoning behind the prediction decision which a highly accurate model may not. Interpretable models give important insights into data and model behaviours and may persuade end-users to use certain models [26]. For users in marketing, medical analysis and science the understanding of data is more important than just predictive accuracy [26]. Being able to explain the reasons behind a decision and validate it is essential in these domains. Our study showed that regressors and clustering decision trees are popular in the latest proposals, with few utilising kNN, neural net and pure clustering. Among the most understandable models are decision trees and decision rules [27]. Neural nets, on the other hand, are widely perceived as black-box models due to their complexity, leaving data scientists little room to explain the resulting prediction [27]. The demanding nature of kNN on resources may have pushed it to the bottom of the selection list.

Table 3. Prediction models

Publications	Prediction models
Adıyeke and Baydoğan [13]	Decision tree
Bessadok et al. [22]	Clustering
Breskvar and Džeroski [19]	Clustering decision tree
Chen et al. [4]	Regressors
Liu et al. [23]	Neural net
Liu et al. [25]	kNN
Mastelini et al. [2]	Regressors
Mignone et al. [24]	Clustering decision trees
Pliakos and Vens [20]	Bi-clustering decision trees
Santana et al. [5]	Regressors
Wu and Lian [21]	Regressors

RQ3: Which MTP strategy that the side information was used in?

Three frequent strategies were found in this study: ensemble, stacking and chaining. Not all publications fall solely in one of these; some are a hybrid of two or more. In general, an ensemble strategy consists of multiple weak prediction models that, when combined, are expected to produce improved performance. Ensembles are known to reduce prediction variance and are resistant to outliers and noisy data [28]. The high execution time of ensembles, on the other hand, is its disadvantage. When using an ensemble, the final performance is usually the average prediction of each run model. In contrast, stacking strategy leverages varying strong predictive models [28]. Models can be regressors, decision trees and even ensembles. The stacking process involves a minimum of two stages of prediction where the prediction from the earlier stage will be used in the later stage to produce better prediction [29]. Stacking also suffers from high execution time as ensembles. A chaining strategy, in general, is where prediction models are tied together to form a chain with the addition that the prediction of an earlier model becomes a supplementary feature to the successive models in the chain [30]. Another form of chaining involves not only a chain of models but also a chain of target variables [14]. Chaining too can be lengthy to complete.

Table 4. Strategy used

Publications	Strategies
Adıyeke and Baydoğan [13]	Ensemble
Bessadok et al. [22]	Stacking
Breskvar and Džeroski [19]	Ensemble
Chen et al. [4]	Stacking
Liu et al. [23]	Stacking
Liu et al. [25]	Ensemble
Mastelini et al. [2]	Stacking
Mignone et al. [24]	Single run
Pliakos and Vens [20]	Ensemble
Santana et al. [5]	Stacking
Wu and Lian [21]	Chaining

RQ4: What metrics were used to evaluate the MTP proposals?

In Table 5, we list the metrics used by each proposal to evaluate the performance of their strategy. We discovered that many of the proposals employ metrics based on the rate of error. The most used are Mean Absolute Error and Root Mean Squared Error or variants of these measures. Under-the-curve metrics could also be found used by binary-based proposals. Other forms of performance measurements include coefficients, centrality metrics, speed of runs and accuracy.

Table 5. Evaluation metrics

Publications	Metrics
Adıyeke and Baydoğan [13]	Relative Root Mean Squared Error.
Bessadok et al. [22]	Pearson Correlation Coefficient. Mean Absolute Error. Betweenness Centrality. Closeness Centrality. Eigenvector Centrality.
Breskvar and Džeroski [19]	Average Relative Root Mean Squared Error.
Chen et al. [4]	Average Coefficient of Determination. Average Root Mean Squared Error. Average Relative Root Mean Squared Error.
Liu et al. [23]	Root Mean Squared Error. Mean Absolute Error.
Liu et al. [25]	Area Under the Precision-Recall curve.
Mastelini et al. [2]	Relative Root Mean Squared Error. Average Relative Root Mean Squared Error. Relative Performance. Runtime.
Mignone et al. [24]	Recall@k. Area Under the Recall@k curve. Area Under the ROC curve. Area Under the Precision-Recall curve.
Pliakos and Vens [20]	Area Under the ROC curve. Area Under the Precision-Recall curve.
Santana et al. [5]	Root Mean Squared Error. Relative Performance per Target. Average Relative Root Mean Squared Error. Ratio of Performance to Deviation.
Wu and Lian [21]	Root Mean Squared Logarithmic Error. Accuracy Rate. Root Mean Squared Error. Mean Absolute Error.

1.2. Depression, Anxiety and Stress (DAS)

Depression, anxiety and stress (DAS) are types of mental illnesses that disable their sufferers in many aspects of life. From having explosive social interactions to performing bodily harm, DAS typically goes unnoticed until it is too late. Typically, clinicians use psychometric scales to determine the severity of a patient's DAS before face-to-face diagnosis similar to ultrasounds done to assist nephrologists. There are a number of psychometric scales dealing with DAS besides DASS-21. They include Hamilton Rating Scale for Depression (HAMD) [31], Montgomery–Åsberg Depression Rating Scale (MADRS) [32], Hospital Anxiety and Depression Scale (HADS) [33], Edinburgh Postnatal Depression Scale (EPDS) [34] and Geriatric Depression Scale (GDS) [35]. Each of these scales measures depression for a specific group of people. We chose DASS-21 as it is a self-reporting psychometric scale designed to measure together with the severity levels of depression, anxiety as well as stress in sufferers. This makes DASS-21 suitable to tap into the mental well-being of the masses since the condition of depression, anxiety and stress are typically interrelated and does not occur in silos. Also, due to the structured nature of DASS-21, conducting them online does not compromise its results. Online here refers to an electronic survey although an earlier study has shown that patients also profess their condition through another online media, that is social media [36]. With the online DASS-21 survey, individuals are to recollect their psychological conditions over a previous week and answer 21 questions regarding it. Scoring will then be calculated to reflect the present mental wellness of the individual. DASS-21 has five levels of severity: Extremely Severe, Severe, Moderate, Mild and Normal. Based on our co-author, Prof. Ramli Musa, psychiatric help is necessary for individuals experiencing Extremely Severe and Severe levels.

The ability to assess DAS based on its severity has long been underscored by the National Institute for Health and Clinical Excellence (NICE) [37] since 2004 for both primary and secondary care [38]. At that time, three severity levels were proposed (mild, moderate and severe) to reflect the escalating symptom count [38]. Now, additions to the levels could be found in the effort to better describe the condition of DAS. The premise is the measurement of severity is the key driver in the determination of suitable psychiatric interventions [37]. Different severity requires different interventions. Without knowing the extent of the severity, psychiatrists are in the dark to work out a wellness plan for patients. Hence, validated assessment tools are imperative and several has been endorsed. They include Patient Health Questionnaire (PHQ-9) [39], Hospital Anxiety and Depression Scale (HADS) and the second edition of the Beck Depression Inventory (BDI-II) [40]. While DASS-21 was not tested in the said validation, however, our comparative study [41] conducted against HADS to examine their concurrent validity showed that DASS-21 not only equally perform well, but excel in the measurement of stress in patients due to the absence of this specific component in HADS. Therefore, DASS-21 is a good tool for assessing severity.

Interest in predicting severity in mental illness can also be seen to escalate as of late with the Covid-19 pandemic sweeping across the globe, triggering multiple crises e.g. financial and psychological. The prospect of being able to anticipate a patient's condition and accordingly deal with it is greatly useful. Now, severity prediction of mental illness has become a repeated component in the eRisk CLEF Workshop [49-50]. The CLEF Initiative (Conference and Labs of the Evaluation Forum) is held annually by a self-organised body seeking to encourage efforts on multilingual and multimodal information in the form of innovation, research and development since 2010. In recent 2019, CLEF has begun introducing eRisk, an effort to perform early detection of depression in online posts, specifically Reddit. This shows that DAS is not a trivial condition, thus must be carefully addressed.

1.3. Our Contributions

We made the following contributions:

1. Introduce a cost matrix as side information for MTP relaying target adjacency knowledge that is validated by a domain expert.
2. Conducted an empirical study on the prediction performance when this cost matrix is used to understand its potential in improving performance.

This paper is structured as follows. In Section 2 we discuss the research method. Our results are described in Section 3. We conclude in the final section.

2. RESEARCH METHOD

In this section, we describe our novel dataset, our proposed side information, our justification of using precision as the measurement of performance and finally, how the empirical study was conducted.

2.1. Proprietary Dataset

We received 270,322 responses from people located in Malaysia and from ethnicities commonly residing in Malaysia i.e., Malay, Chinese and Indian with only a fraction of the respondents falling under the Others category. Demographic information was recorded while respondents' identities remain anonymous. Both English and Malay language versions of the DASS-21 were publicly accessible on our website . The website may be recognized by respondents through informal means such as word of mouths, social media and blogs. Respondents were not invited to partake in the survey but instead were random people who seeks an understanding of their condition, thus willingly taking the DASS-21 survey. Complete anonymity was given to all respondents where no information of name or email address were collected. The attributes collected are marital status, gender, age, race, education level and occupation. A severity score is calculated from the 21 questions asked in DASS (Malay and English versions). The Malay version is known as MDASS-21. Respondents will immediately receive their severity scoring upon completion. Table 6 presents the size and details of the dataset and Figure 1 visualizes the details. To predict severity, we used the assigned severity levels as the labels. The dataset is imbalanced, predominated by Extremely Severe and Moderate data points. 859 data were found to be unfit due to missing values of gender and race, hence, have been removed. None of the responses was Normal across all depression, anxiety and stress classes. The dataset was automatically annotated with the severity levels from the DASS-21 survey. Careful inspection found the absence of negative class as every severity level, inclusive of Normal, are interesting classes. This is justified by the same importance of predicting the Normal severity level to other levels.

Table 6. Dataset

Normal	23,139
Mild	35,522
Moderate	74,087
Severe	45,588
Extremely Severe	91,127
Dirty	859
Total	**270,322**

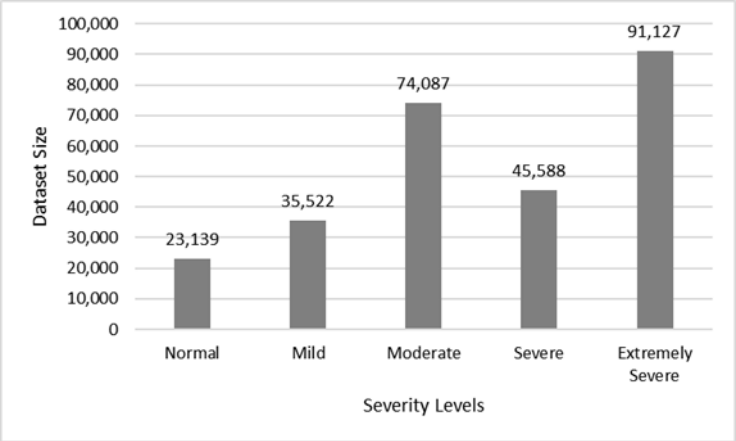


Figure 1. Distribution of dataset

2.2. Adjacency Side Information

The side information that we study in this paper is of targets that are highly correlated and there exists an internal structure of adjacency that could be implemented on the targets to improve performance. Adjacency reflects the "alikeness" of a particular target outcome to another target outcome. The element of alikeness can be fed into the base models to enhance performance. In the case of predicting depression, anxiety and stress, the levels of severity can be found at differing degrees of extremely severe, severe, moderately severe, mildly severe and normal. The alikeness of a moderately severe condition is closer to severe and mildly severe than normal and extremely severe. Alikeness here refers to the faded line between the adjacent levels in describing a sufferer's degree of severity. To that end, we apply a matrix to represent the adjacency information (Table 7). A value of 0.0 denotes the same severity level, and as two severity levels get further apart, a higher value is assigned. A value of 4.0 represents the greatest degree of unalikeness.

Table 7. Adjacency matrix

	Normal	Mild	Moderate	Severe	Extremely Severe
Normal	0.0	1.0	2.0	3.0	4.0
Mild	1.0	0.0	1.0	2.0	3.0
Moderate	2.0	1.0	0.0	1.0	2.0
Severe	3.0	2.0	1.0	0.0	1.0
Extremely Severe	4.0	3.0	2.0	1.0	0.0

For the information to be observed by the base learners, we take the approach of implementing a penalization system, the cost matrix. The cost matrix has been used in cost-sensitive learning to assist in making optimal decisions [12]. Using cost to represent severity is not a new idea – other existing representations include money and waste of time [12]. Our approach to applying adjacency information to the cost matrix is supported by the premise that different levels of severity will require different treatments. An extremely severe case that warrants psychiatric help, if mispredicted to be the complete opposite i.e. normal, can result in being overlooked and thus may result in tragedy, such as loss of life or extreme bodily harm, since it will go untreated. Such devastating outcome also applies to normal cases which, if mispredicted, can eventually turn into depression, anxiety or stress due to false belief. This misprediction is reflected in the cost matrix, with 4.0 being the highest cost. In contrast, a low cost of 1.0 is given to adjacent severity levels for example moderate-mild and moderate-severe. This represents the lesser negative implication that can occur due to incorrect prediction, considering that there is a thin line to discriminate between adjacent levels of severity. As two levels become further apart, the higher the cost of incorrect prediction becomes to indicate the rising criticality of a misprediction. This is shown in Table 8 where mild-severe misprediction score is 2.0 and mild-extremely severe's score is 3.0.

Table 8. Cost matrix

	Actual Normal	Actual Mild	Actual Moderate	Actual Severe	Actual Extremely Severe
Predicted Normal	0.0	1.0	2.0	3.0	4.0
Predicted Mild	1.0	0.0	1.0	2.0	3.0
Predicted Moderate	2.0	1.0	0.0	1.0	2.0
Predicted Severe	3.0	2.0	1.0	0.0	1.0
Predicted Extremely Severe	4.0	3.0	2.0	1.0	0.0

We thus define a cost-sensitive multi-target prediction that incorporates the adjacency information.

Definition (Cost-sensitive multi-target prediction): A cost-sensitive MTP is characterised with a given set of instances $\{x_1, \ldots, x_n\}$ and a set of targets $\{t_1, \ldots, t_n\}$ to be predicted based on a score sensitive to cost values in a matrix $C(p, a)$ where p is the prediction class and a is the actual class. A correct prediction is when $p = a$ yielding a cost of 0, whereas an incorrect prediction is when $p \neq a$ and producing a value $v > 0$. Therefore, a cost-sensitive multi-target prediction can be defined as a triplet of an instance, a target and a cost-sensitive prediction score.

$$M(x) = (x_n, t_n, C(p, a))$$

Figure 2 illustrates our proposed method. The labelled dataset underwent preprocessing and afterwards cross-validation (10 folds). During each fold, training, testing and evaluation processes were conducted to produce a precision score. Training of a machine learning algorithm was done guided by our cost matrix.

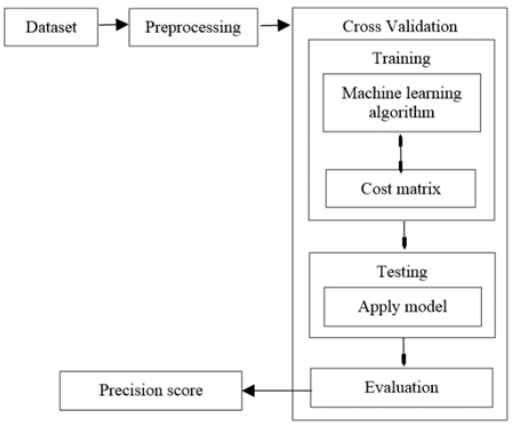


Figure 2. Cost-sensitive MTP

2.3. Multi-Target Prediction Strategy

Based on the review of the latest literature on MTP strategies in Section 1.1, stacking was the most used, therefore, it was selected for this study. Stacking falls under the multiple algorithms group of prediction strategies together with ensemble and chaining (Figure 3). Generally, they produce better performance than a classical single algorithm strategy by leveraging on the power of many. Stacking's advantage is in the use of strong learners to produce a prediction that will be used by another strong learner for improved performance. Ensemble and chaining, on the other hand, are devised to utilize on the unique traits of weak learners. In this strategy, our labelled dataset was divided into two sets: training and testing. The training set was used with a group of base learners to produce a model that was then applied to the testing set to produce a base prediction. Stacking was performed by inputting the resulting base prediction into a stacking learner. The prediction generated from this process was the final classification result. Refer to Figure 4 for the stacking algorithm.

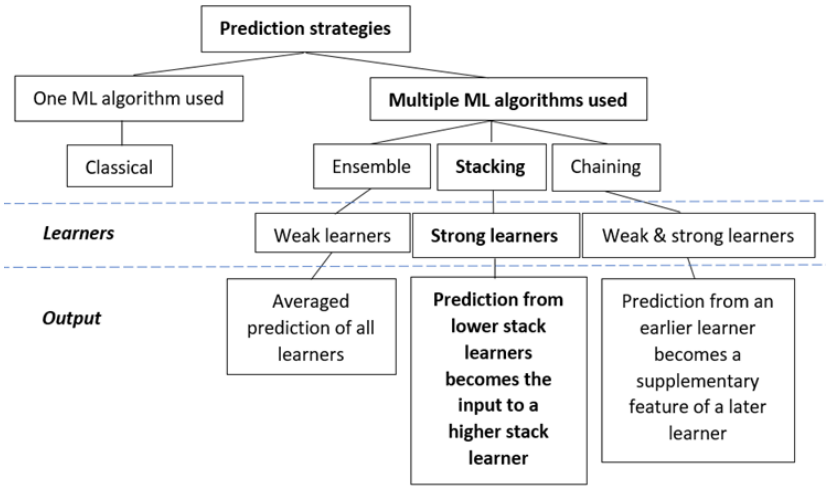


Figure 3. Prediction strategies

Algorithm Stacking
Input *T: A labelled dataset*
B: A set of base learners
S: A stacking learner
Output *P: Performance*
Begin
Divide T into X and Y sets
Model = Train(B, X)
BPrediction = Test(B, Y, Model)
SPrediction = Train(BPrediction, S)
End

Figure 4. General stacking algorithm

2.4. Evaluation Metrics

Metrics used to evaluate the performance of the adjacency information were averaged accuracy (AA), mean recall (MR), mean precision (MP) and averaged root mean squared error (ARMSE). The multi-label and multi-class nature of the dataset necessitates the use of averaging to conclude a single final value of evaluation. Here, no weights were employed to calculate MR and MP since our aim was to study the effects of the proposed adjacency information injected during the training of the models. Therefore, any penalization was learned in the model induction process. A tradeoff between metrics is inevitable, thus, in this study MP and MR takes precedence over AA. MP answers the question of "*What is the averaged proportion of positive identifications was correct?"* while MR sets to know "*What proportion of actual positives was identified correctly?*" – underlining critical aspects of training a model to incline towards correctly predicting positive classes more than negative classes. This is well suited to our proprietary dataset as well as other datasets inherent with the same characteristics. On the other hand, AA sought to answer "*What is the averaged proportion of correct predictions (positive and negative classes) over the entire dataset?*". Hence, a higher MP and MR is more preferred in this particular case than AA. ARMSE indicates the percentage of error that the algorithms produced, hence, preferably as low as possible.

2.5. Empirical Study

Our empirical study aims to learn if our adjacency information can contribute to the improvement of performance when applied to existing machine learning algorithms. This study should lead us to future works on suggesting more use of the adjacency information, other improved forms of the adjacency information and strategies specialized to the adjacency information. We conducted three empirical studies. The algorithms used were decision tree (DT), Naïve Bayes (NB), generalized linear (GL) and deep learning (DL). During testing, 10 folds cross-validation were used for each algorithm, yielding an averaged performance value. The choice of the number of folds was based on common practice. The cost matrix is constructed based on the cost information displayed in Table 8.

Experiment 1: Adjacency vs. None (Figure 5)

The aim is to know the hidden potential of the adjacency information in improving performance when employed to existing algorithms. This would also determine its direct use with these algorithms, or if a specially constructed algorithm is necessary. No baseline algorithm was chosen as the focus is on discovering which algorithm responds most positively and be used as the stacking model in Experiment 2. To this end, the precision generated with and without the adjacency information is compared.

Result: None of the algorithms achieved the 50% mark on AA, WMP and WMR. Expectedly, the performance when no adjacency information is used will be higher for AA and WMR since no penalizations were conducted when mispredictions occur. Nevertheless, DL, GL and NB increased in WMP when adjacency information is used but this is not true for DT. WMP provides the average of correct predictions of positive classes. This is very relevant to this study as the dataset consists more of this class and the focus is to correctly predict it. The result shows DT is not suitable for this type of side information, recording the lowest WMP (9.79). NB yield the highest value (30.07) for WMP, hence, is the most suitable when working with adjacency information. This is followed by GL (28.45) and DL (27.08). With none reaching 50% of performance, this shows a gap exists to be filled. ARMSE for NB, GL and DL do not differ considerably with or without adjacency information employed, but is highest for DT, supporting its unsuitability for adjacency information. Due to space limitations, this paper only displays selected confusion matrices as presented in Table 9 till 12.

Table 9. Confusion matrix for NB without adjacency information

		Actual					
		Extremely Severe	**Severe**	**Normal**	**Moderate**	**Mild**	**Precision**
Prediction	**Extremely Severe**	48097	19263	21674	29622	9069	37.66%
	Severe	31	19	41	37	9	13.87%
	Normal	23373	18957	45662	32276	17011	33.26%
	Moderate	11934	11437	21655	20459	16418	24.98%
	Mild	4871	3181	4841	7026	10285	34.05%
	Recall	54.47%	0.04%	48.64%	22.88%	19.48%	

Table 10. Confusion matrix for DT without adjacency information

		Actual					
		Extremely Severe	**Severe**	**Normal**	**Moderate**	**Mild**	**Precision**
Prediction	**Extremely Severe**	23681	14922	4766	11342	9620	36.81%
	Severe	2650	5006	3244	2486	1610	33.38%
	Normal	988	1452	1611	1523	836	25.13%
	Moderate	10797	14167	16128	24297	10394	32.06%
	Mild	36	32	22	36	32	20.25%
	Recall	62.07%	14.07%	6.25%	61.23%	0.14%	

Table 11. Confusion matrix for NB with adjacency information

		Actual					
		Extremely Severe	**Severe**	**Normal**	**Moderate**	**Mild**	**Precision**
Prediction	**Extremely Severe**	9486	3683	3991	6407	6544	31.50%
	Severe	35302	17408	25300	26939	10061	15.14%
	Normal	1686	2897	10433	4101	1563	50.45%
	Moderate	39847	27524	50204	49111	33499	24.53%
	Mild	1985	1345	3945	2862	1125	9.99%
	Recall	10.74%	32.93%	11.11%	54.92%	2.13%	

Table 12. Confusion matrix for DT with adjacency information

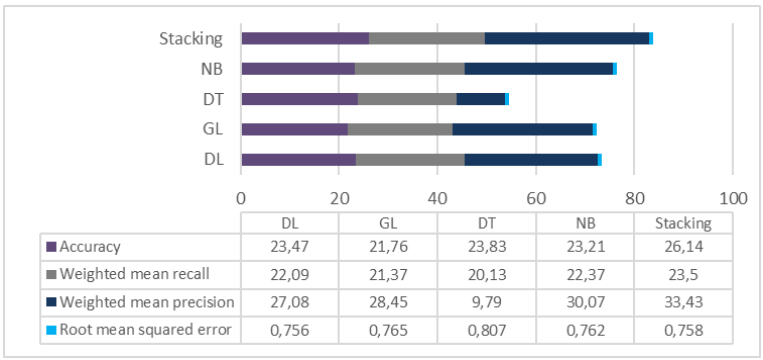
		Actual					
		Extremely Severe	**Severe**	**Normal**	**Moderate**	**Mild**	**Precision**
Prediction	**Extremely Severe**	11	34	11	6	6	16.18%
	Severe	27002	19005	6229	16538	11648	23.63%
	Normal	10808	16387	19123	21662	10452	24.38%
	Moderate	0	0	0	0	0	0.00%
	Mild	331	153	408	1478	386	14.01%
	Recall	0.03%	53.42%	74.20%	0.00%	1.72%	

Figure 5. Adjacency vs. None

Experiment 2: Stacking vs. No stacking (Figure 6)

The aim is to understand the adjacency information's applicability to both predictive strategies in producing a good performance. The outcome can help researchers to decide on the more suitable strategy when dealing with the proposed adjacency information. Here, stacking was used to represent the multiple algorithms strategy. From Experiment 1, NB was selected as the stacking learner while GL and DL as base learners; DT is excluded due to its unsatisfying performance. On this account, each performance metric of individual learners is contrasted against stacking them together.

Result: The stacking approach churned the highest ACC, WMR and WMP results. Besides, it resulted in low ARMSE. Overall, stacking performed better than the learners running individually on all fronts. This indicates using a stacking design of the strongest learner as the stacking learner and lesser strong learners as base learners, simultaneously removing unsatisfying learners helps to improve performance. Devising a different stacking design may extend performance, hence, an open research opportunity. Also, we hypothesize tweaking the penalization system can further enhance performance. Penalization system differs from domain to domain. In this study, the penalization system relies on the distance between classes to train the underlying algorithms and is quite strict. Relaxing this or focusing on a different aspect of penalization besides distance may boost performance.



	DL	GL	DT	NB	Stacking
Accuracy	23,47	21,76	23,83	23,21	26,14
Weighted mean recall	22,09	21,37	20,13	22,37	23,5
Weighted mean precision	27,08	28,45	9,79	30,07	33,43
Root mean squared error	0,756	0,765	0,807	0,762	0,758

Figure 6. Stacking vs. No Stacking

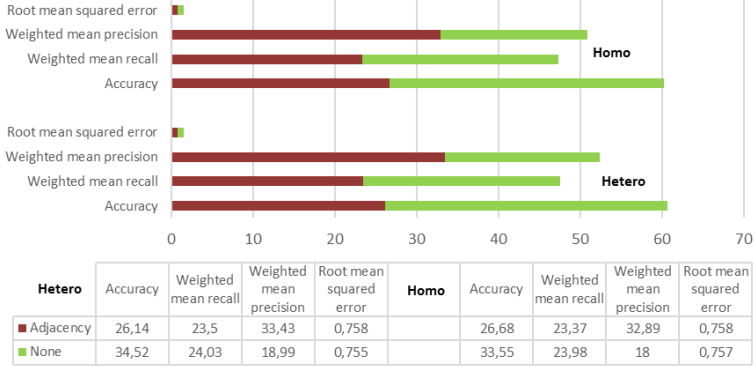
Experiment 3: Heterogeneous vs. Homogeneous models (Figure 7)

The aim is to ascertain if the adjacency information will respond positively to stacking with either homogeneous or heterogeneous models. A homogeneous model consists of a single algorithm run in multiple instances whereas a heterogeneous model has numerous algorithms jointly stacked. We used the stacking design proposed in Experiment 2 to test for a heterogeneous model. NB, being the best algorithm, was chosen for the homogeneous model. We also tested the performance of these two models when adjacency information is present.

Result: Heterogeneous model obtained better results than the homogeneous model with or without adjacency information, except for AA where the heterogeneous model achieved 26.14 and the homogeneous model superseded at 26.68. In this predominantly positive class dataset, leveraging on the strength of many has resulted in more mispredictions during converging. Nevertheless, NB is known to score well in accuracy. Thus, the homogeneous model does not inherit the weaknesses of the other algorithms. Predominantly, all ARMSE values produced are closely similar, indicating the comparable ability of both models. Reducing ARMSE is a possible research focus.

3. CONCLUSION

In conclusion, improving performance in multi-target prediction using adjacency information is promising. Predicting DAS' severity level is a domain that could benefit from this notion. In this work, we have studied side information for MTP in the form of a cost matrix that penalizes incorrect prediction of severity levels regarding DAS based on expert knowledge. We have also collected a proprietary labelled dataset of severity levels on depression, anxiety and stress based on DASS-21 from 270,322 respondents in Malaysia. Additionally, we have defined a cost-sensitive multi-target prediction method and finally, we have conducted an empirical study on the prediction performance when adjacency information in the form of cost matrix is used. We discovered that the adjacency information can be used with existing machine learning algorithms and a study to further improve the performance is necessary. The best performing algorithm to be used with this adjacency information is NB and the worst is DT. Also, a stacking strategy of heterogeneous algorithms can obtain better performance compared to individual algorithms when it was designed to comprise of different types of strong learners and exclude unsatisfying learners. The area of side information in MTP is bound to grow further with more new applications and uses are found.



Hetero	Accuracy	Weighted mean recall	Weighted mean precision	Root mean squared error	Homo	Accuracy	Weighted mean recall	Weighted mean precision	Root mean squared error
Adjacency	26,14	23,5	33,43	0,758		26,68	23,37	32,89	0,758
None	34,52	24,03	18,99	0,755		33,55	23,98	18	0,757

Figure 7. Heterogeneous vs. Homogeneous models

ACKNOWLEDGEMENTS

We are extremely grateful to the reviewers who took the time to provide constructive feedback and useful suggestions for improving this article. The Malaysian government is funding this research through the Fundamental Research Grant Scheme (FRGS) at Universiti Teknologi MARA (UiTM) Shah Alam, Malaysia (FRGS/1/2019/SS05/UITM/02/5).

REFERENCES

[1] W. Waegeman, K. Dembczyński, and E. Hüllermeier, "Multi-target prediction: a unifying view on problems and methods," *Data Mining and Knowledge Discovery*, vol. 33, no. 2, pp. 293–324, 2019, doi: 10.1007/s10618-018-0595-5.

[2] S. M. Mastelini, E. J. Santana, R. Cerri, and S. Barbon Jr, "DSTARS: a multi-target deep structure for tracking asynchronous regressor stacking," *Applied Soft Computing*, vol. 91, p. 106215, 2020.

[3] L. Masera, "Multi-target Prediction Methods for Bioinformatics: Approaches for Protein Function Prediction and Candidate Discovery for Gene Regulatory Network Expansion," 2019.

[4] S. Chen, C. Gu, C. Lin, and M. A. Hariri-Ardebili, "Prediction of arch dam deformation via correlated multi-target stacking," *Applied Mathematical Modelling*, vol. 91, pp. 1175–1193, 2021.

[5] E. J. Santana, F. R. dos Santos, S. M. Mastelini, F. L. Melquiades, and S. Barbon Jr, "Improved prediction of soil properties with multi-target stacked generalisation on EDXRF spectra," *Chemometrics and Intelligent Laboratory Systems*, vol. 209, p. 104231, 2021.

[6] X. Wang, X. Zhen, Q. Li, D. Shen, and H. Huang, "Cognitive assessment prediction in Alzheimer's disease by multi-layer multi-target regression," *Neuroinformatics*, vol. 16, no. 3, pp. 285–294, 2018.

[7] S. B. Junior *et al.*, "Multi-target prediction of wheat flour quality parameters with near infrared spectroscopy," *Information processing in agriculture*, vol. 7, no. 2, pp. 342–354, 2020.

[8] E. Gomede, R. de Barros, and L. de Souza Mendes, "Use of Deep Multi-Target Prediction to Identify Learning Styles," *Applied Sciences*, vol. 10, no. 5, p. 1756, 2020.

[9] F. Adilova and A. Ikramov, "Using Support Vector Regression in multi-target prediction of drug toxicity," in *2020 IEEE 14th International Conference on Application of Information and Communication Technologies (AICT)*, 2020, pp. 1–3.

[10] S. G. Fashoto, A. S. Metfula, B. B. Matsebula, and B. Y. Fashoto, "Multi-Target Regression Prediction on Cervical Cancer for Evaluation of Predictive Performance Measures," *Asian Journal of Information Technology*, vol. 17, no. 2, pp. 160–166, 2018.

[11] J. Palmer, V. S. Sheng, T. Atkison, and B. Chen, "Classification on grade, price, and region with multi-label and multi-target methods in wineinformatics," *Big Data Mining and Analytics*, vol. 3, no. 1, pp. 1–12, 2019.

[12] Z. Jing and others, "Multi-Target Prediction Algorithm Based on AdaBoost Regression Tree," *Computer and Modernization*, no. 9, p. 89, 2017.

[13] E. Adıyeke and M. G. Baydoğan, "The benefits of target relations: A comparison of multitask extensions and classifier chains," *Pattern Recognition*, vol. 107, 2020, doi: 10.1016/j.patcog.2020.107507.

[14] G. Melki, A. Cano, V. Kecman, and S. Ventura, "Multi-target support vector regression via correlation regressor chains," *Information Sciences*, vol. 415, pp. 53–69, 2017.

[15] Y. Chen and M. de Rijke, "A collective variational autoencoder for top-n recommendation with side information," in *Proceedings of the 3rd Workshop on Deep Learning for Recommender Systems*, 2018, pp. 3–9.

[16] D. Kang, D. Dhar, and A. B. Chan, "Incorporating side information by adaptive convolution," 2017.

[17] S. Narayan, N. Papasarantopoulos, S. B. Cohen, and M. Lapata, "Neural extractive summarization with side information," *arXiv preprint arXiv:1704.04530*, 2017.

[18] T. Marew, J. Kim, and D. H. Bae, "Systematic Mapping Studies in Software," *International Journal of Software Engineering & Knowledge Engineering*, vol. 17, no. 1, pp. 33–55, 2007, [Online]. Available: http://content.ebscohost.com/ContentServer.asp?T=P&P=AN&K=22674743&S=R&D=bth&EbscoContent=dGJyMNHX8kSeqK44zdnyOLCmr0qeprZSr6e4SrCWxWXS&ContentCustomer=dGJyMPGosk+xq65QuePfgeyx44Dt6fIA%5Cnhttp://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=2447601.

[19] M. Breskvar and S. Džeroski, "Multi-Target Regression Rules With Random Output Selections," *IEEE Access*, vol. 9, pp. 10509–10522, 2021.

[20] K. Pliakos and C. Vens, "Drug-target interaction prediction with tree-ensemble learning and output space reconstruction," *BMC bioinformatics*, vol. 21, no. 1, p. 49, 2020.

[21] Z. Wu and G. Lian, "A novel dynamically adjusted regressor chain for taxi demand prediction," *Proceedings of the International Joint Conference on Neural Networks*, 2020, doi: 10.1109/IJCNN48605.2020.9207160.

[22] A. Bessadok, M. A. Mahjoub, and I. Rekik, "Topology-Aware Generative Adversarial Network for Joint Prediction of Multiple Brain Graphs from a Single Brain Graph," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12267 LNCS, pp. 551–561, 2020, doi: 10.1007/978-3-030-59728-3_54.

[23] F. Liu, Y. Lu, and M. Cai, "A hybrid method with adaptive sub-series clustering and attention-based stacked residual LSTMs for multivariate time series forecasting," *IEEE Access*, vol. 8, pp. 62423–62438, 2020, doi: 10.1109/ACCESS.2020.2981506.

[24] P. Mignone, G. Pio, S. Džeroski, and M. Ceci, "Multi-task learning for the simultaneous reconstruction of the human and mouse gene regulatory networks," *Scientific Reports*, vol. 10, no. 1, Dec. 2020, doi: 10.1038/s41598-

020-78033-7.
[25] B. Liu, K. Pliakos, C. Vens, and G. Tsoumakas, "Local Imbalance based Ensemble for Predicting Interactions between Novel Drugs and Targets," *PharML 2020 Machine Learning for Pharma and Healthcare Applications*, pp. 1–5, 2020.
[26] J. Wang, R. Fujimaki, and Y. Motohashi, "Trading interpretability for accuracy: Oblique treed sparse additive models," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 2015-Augus, pp. 1245–1254, 2015, doi: 10.1145/2783258.2783407.
[27] M. Breskvar and S. Džeroski, "Multi-Target Regression Rules With Random Output Selections," *IEEE Access*, vol. 9, pp. 10509–10522, 2021.
[28] B. Boehmke and M. G. Brandon, *Hands-On Machine Learning with R*. CRC Press, 2019.
[29] E. J. M. Lauría, E. Presutti, M. Kapogiannis, and A. Kamath, "Stacking classifiers for early detection of students at risk," *CSEDU 2018 - Proceedings of the 10th International Conference on Computer Supported Education*, vol. 1, no. Csedu 2018, pp. 390–397, 2018, doi: 10.5220/0006781203900397.
[30] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains: A review and perspectives," *Journal of Artificial Intelligence Research*, vol. 70, pp. 683–718, 2021, doi: 10.1613/JAIR.1.12376.
[31] M. Hamilton, "The Hamilton Rating Scale for Depression," in *Assessment of Depression*, 1986, pp. 143–152.
[32] S. A. Montgomery and M. Åsberg, "A new depression scale designed to be sensitive to change," *The British journal of psychiatry*, vol. 134, no. 4, pp. 382–389, 1979.
[33] R. P. Snaith, "The hospital anxiety and depression scale," *Health and quality of life outcomes*, vol. 1, no. 1, pp. 1–4, 2003.
[34] J. L. Benvenuti, P., Ferrara, M., Niccolai, C., Valoriani, V., & Cox, "Detection of postnatal depression: development of the 10-item Edinburgh Postnatal Depression Scale," *Journal of affective disorders*, vol. 53(2), pp. 137–141, 1999.
[35] J. A. Sheikh, J. I., & Yesavage, "Geriatric Depression Scale (GDS): recent evidence and development of a shorter version," *Clinical Gerontologist: The Journal of Aging and Mental Health*, 1986.
[36] M. Z. Nasrudin, R. Maskat, and R. Musa, "Detecting candidates of depression, anxiety and stress through Malay-written tweets: A preliminary study," *Indonesian Journal of Electrical Engineering and Computer Science*, 2019, doi: 10.11591/ijeecs.v16.i2.pp787-793.
[37] "National Institute for Clinical Excellence. Depression: management of depression in primary and secondary care (NICE guideline)." London: National Institute for Clinical Excellence, 2004.
[38] I. M. Cameron, J. R. Crawford, K. Lawton, and I. C. Reid, "Psychometric comparison of PHQ-9 and HADS for measuring depression severity in primary care," *British Journal of General Practice*, vol. 58, no. 546, pp. 32–36, 2008, doi: 10.3399/bjgp08X263794.
[39] R. L. Kroenke, K., & Spitzer, "The PHQ-9: a new depression diagnostic and severity measure," *Psychiatric annals.*, vol. 32(9), pp. 509–515, 2002.
[40] G. Beck, A. T., Steer, R. A., & Brown, "Beck Depression Inventory," *Psychological Assessment*.
[41] R. Musa, R. Ramli, K. Abdullah, and R. Sarkarsi, "Concurrent validity of the depression and anxiety components in the Bahasa Malaysia version of the Depression Anxiety and Stress Scale (DASS)," *ASEAN Journal of Psychiatry*, vol. 12, no. 1, p. Jan-June, 2011.
[42] D. E. Losada, F. Crestani, and J. Parapar, "eRISK 2017: CLEF lab on early risk prediction on the internet: experimental foundations," in *International Conference of the Cross-Language Evaluation Forum for European Languages*, 2017, pp. 346–360.
[43] D. E. Losada, F. Crestani, and J. Parapar, "Overview of eRisk at CLEF 2019 Early Risk Prediction on the Internet (extended overview)," *CEUR Workshop Proceedings*, vol. 2380, no. September, pp. 9–12, 2019.
[44] R. Jonschkowski, S. Höfer, and O. Brock, "Patterns for Learning with Side Information," 2015, [Online]. Available: http://arxiv.org/abs/1511.06429.