Journal of Theoretical and Applied Information Technology
15th October 2020. Vol.98. No 19

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

SYMPTOM-BASED DISEASE PREDICTION SYSTEM USING MACHINE LEARNING

[1] JING YI LEONG, [2] BOOMA P M

[1] Undergraduate Student, School of Computing, Engineering & Technology, Asia Pacific University of Technology & Innovation (APU), Kuala Lumpur, Malaysia.
[2] Lecturer, School of Computing, Asia Pacific University of Technology & Innovation (APU), Kuala Lumpur, Malaysia.
Email: [1] jingyileonggg@gmail.com, [2] dr.booma@staffemail.apu.edu.my

ABSTRACT

Healthcare has been an important industry from then till now, and it is said to be one of the sectors which plays a critical role in preventing the increasing number of a particular disease. In this era of new technology, machine learning has been used in a lot of industries, and one would be the healthcare industry. In the healthcare field, machine learning contributes significantly to predicting a disease as to simplify the process of the manual disease diagnosis and bring convenience to both the doctor and patient. In this paper, a disease prediction system will be implemented with the use of supervised learning algorithm to allow patient in identifying disease themselves based on their symptoms. Few supervised learning algorithms are being trained and tested in terms of their accuracy, and the algorithm with the highest accuracy is used for the prediction. The chosen supervised learning algorithms to be tested include Bernoulli Naïve Bayes, Decision Tree, and Support Vector Machine.
Keywords: *Classification, Disease, Healthcare, Machine Learning, Prediction*.

1. INTRODUCTION

Having a doctor or pharmacist consultation is a must when getting sick. Generally, doctor is one of the professions in healthcare field, who is responsible for both general and specialty disease diagnosis [1]. However, there is a problem where patients may not access to healthcare services when they met some common clinical diseases such as dengue, chickenpox, allergy, and so forth. Transportation issue, the amount spent on medical services, as well as the low health awareness on patients themselves may be the major issues that cause the problems mentioned to have happened. Other than pharmacist and doctor, people are facing a challenge where he or she does not have any experiences when coming to the diagnosis of clinical diseases such as dengue, chickenpox, allergy, and so forth.

In the 21st century era of technology, machine learning has been evolved and is widely used in various industries, including agriculture, finance, travel, and so forth [2]. One of the significant fields which involves machine learning is the medical field [3]. Different kinds of machine learning approaches have been used in the medical industry, such as natural language processing to create a medical diagnosis chatbot, image processing to determine rare disease, an expert system to determine general disease, and so on [4]. Since the project relates to prediction in the healthcare field, therefore machine learning will be used for prediction in the project.

Based on the problems stated in the first paragraph, an offline-based medical diagnosis system is proposed to help patients in identifying diseases themselves. After selecting symptoms, they will receive an accurate diagnosis result from the system. In short, patients can know the disease they have and later decide on whether there is a need to go for doctor consultation.

In order to produce an accurate diagnosis result, a set of data will be trained by using appropriate machine learning algorithms. After testing part of the data for its accuracy, data (symptoms) entered by the patient will be input and the outcome will be produced based on the data trained using machine learning algorithms. The dataset collected would be

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

the medical diagnosis records of other patients based on their symptoms.

2. PROBLEM CONTEXT

Consulting a doctor may cause someone to have huge spending on it. According to the research found, the healthcare cost of the United States (US) has risen from $2879 billion to $3492.10 billion in five-year time (from the year of 2013 to 2017) [5]. Also, according to Deloitte (2019), the spending of worldwide medical care is presumed to be increased continuously from the year of 2017 to 2022, with an annual rate of 5.4% [6]. The rising cost in the medical field causes part of the people not able to pay for the clinical or hospital consultation, especially for the lower-income group, and this causes more people choose to not seek for medical care when they are having clinical diseases such as allergy, high fever, and so forth [7]. When diseases getting worsen, the chance of getting effective treatment may be lower.

There is a problem where people nowadays have low health awareness on themselves [8]. As a worker, he or she may not willing to consult a doctor when getting sick, as there are tons of jobs needed to be done by them every day [9]. Therefore, health awareness of most of the people become lower, which may cause the disease to become more serious in the future. Besides, as mentioned in the first paragraph, consulting a doctor may need to spend a lot on it. This may lead to the same problem where health awareness becomes lower and causing the disease to get more serious.

Going to a physical clinic or hospital for doctor consultation may be hard for people who live in rural areas [10]. According to research, the number of medical clinics or hospitals in rural areas is much lesser compared to in urban areas [11]. This will lead to a problem where people who live in rural areas have fewer chances in obtaining medical care they need [12]. If rural residents want to access a particular medical service, they may need to have their transportation to reach the medical service, in which the medical service may locate far from home [13].

3. AIM

To come out with an implementation of a symptom-based disease prediction system using machine learning algorithm to bring convenience to users by identifying the clinical disease themselves based on a set of past data.

4. OBJECTIVES

The project objectives include:

- To have an investigation of few similar existing systems.
- To identify the problems of each similar existing system so that these problems can be improved within the proposed system.
- To compare various machine learning algorithms and identify the most suitable algorithm for prediction.
- To evaluate each chosen algorithm in terms of its accuracy and choose the best one to be implemented within the system.

5. LITERATURE SURVEY

Doctor is one of the professions in the healthcare field among the others such as nurse, dentist, psychologist, and so forth. The doctor-patient relationship can be defined as when someone is getting sick, he or she has to consult a doctor in either clinic or hospital, and the doctor has the responsibility to help patients in diagnosing their disease based on symptoms and providing suggestions and medication to them [1]. However, there are still some people who may have hesitation before consulting a doctor, and this may due to multiple reasons as stated in the problem context, such as cost, time, as well as the accessibility of medical service. Without treating the sickness immediately, a minor illness may have the chance of transforming into a major illness at the end of the day [14]. Therefore, the implementation of an offline-based disease prediction system will be carried out to act as a solution in solving these problems.

5.1. Background of the Healthcare Industry

According to the research, healthcare service acts as the biggest and one of the most important industries in the United States [15]. Various treatments have been offered to patients through the access to medical services, and patients are able to get consultations from healthcare professions such as pharmacist and doctor [16]. According to one of the reports found, the worldwide healthcare market is growing continuously by reaching a value of approximately $8452 billion in the year of 2018, and it is expected to grow to approximately $11908.9 billion in the year of 2022 [17]. From here, a conclusion can be made that healthcare services may be one of the industries that can be further explored and improved with the existence of technologies in

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

this 21st century, which is the era of science and technology.

However, there is still part of the people who choose to not accept treatment from the healthcare profession and their health condition is deteriorating at the end [18]. Therefore, it is assumed that these people may have a few concerns before getting treatment. A study was carried out to discover the reasons why patients chose to avoid from accessing to medical service [16]. From the study, it can be discovered that there are actually multiple aspects which make the respondents to avoid from getting themselves a medical care, including interpersonal factor such as communication issues, traditional barrier such as transportation difficulties, as well as organizational factor such as long waiting time. To solve these problems, an in-depth research is carried out by exploring whether or not the evolvement of technology in the medical field can help in solving the problems stated.

As mentioned above, technology has been quietly and slowly evolved in the healthcare industry in this era of new technology [19]. One of the reports mentioned that the market growth of smart healthcare industry is expected to reach a CAGR of 24.1% between the year of 2019 and 2023 [20]. From here, it can be summarized that the medical industry will continuously adopt smart technology now and in the future. Smart healthcare industry can be said as the combination of either big data, internet of things (IoT) or machine learning, as well as deep learning with the healthcare system [21]. Using a smart healthcare system enables the patients to not only have the option of accessing physical hospital, but also the option to access to the healthcare system anywhere at any time without any boundaries [22]. From here, it can be known that both issues (transportation difficulties and long waiting times) can be solved through the smart healthcare system, as it allows patients to access to healthcare service anytime at anywhere. Besides, the smart healthcare system allows the cost of medical to be reduced as well [23]. All in all, it can be concluded that the utilization of technology in healthcare industry is able to solve all the problem contexts mentioned.

5.2. Manual Doctor Consultation Process

Due to privacy issue, secondary research is carried out instead of primary research to understand the manual process of doctor consultation, so that the proposed system is able to mimic the procedure well, but with a simplified process.

A study was carried out by an author to investigate the time taken for the patient to wait for doctor consultation as well as the time taken for the patient to consult a doctor at a clinic so that the author can produce few strategies to improve this matter [24]. Another research was carried out by an author to improve the old system of doctor consultation by designing an electronic consultation system [25]. From both of these journals, the researcher is able to know and understand the walk-in process has to go through a lot of processes, including registration, pre-consultation, consultation, booking of appointment, payment process, as well as pharmacy process (wait for the medicine to be given by the doctor). From here, it is assumed that this is one of the reasons why some of the people think that they have to spend a lot of time on manual doctor consultation, as there are a lot of processes to be gone through.

5.3. Introduction to Machine Learning

There are a lot of subsets involved in artificial intelligence, and one of them would be machine learning. Machine learning has been widely used in various industries, including medical, agriculture, manufacturing, sales, and so on [26]. By using machine learning, it provides the opportunity for the system to learn from a chunk of past data itself and improve by the experiences automatically [27]. One of the functions of machine learning is that it acts as an assistant within data analytic field [28]. There is various form of data analytics, including descriptive analytic, predictive analytic, as well as prescriptive analytics [29]. In predictive analytics, machine learning algorithm is used to analyze the past data collected and to predict future output [28]. Generally, machine learning allows the system to learn by looking at the data such as the pattern of numeric data, instructions, and so forth [30]. After the data is being learned or trained, it allows the machine to make a better decision in the future. There are three types of machine learning, including unsupervised learning, supervised learning, as well as reinforcement learning [31].

Supervised learning learns through a set of existing data consisting of both input data and output data, in which the output data is the correct answer [32], [33]. The learning process will be terminated when the algorithm reaches a good performance. There are two common groups that fall under supervised learning, including classification and regression [34]. Classification problem produces an output with a category, such as “yes” or “no”; while

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

regression problem produces an output with an exact value, such as "height".

On the other hand, unsupervised learning refers to a machine learning algorithm that aims to search for unknown patterns or structures in the data, from a set of data that has input data but without any labeled output [35]. Using unsupervised learning allows the data to be categorized based on their similarities [36]. There are two common groups that fall under unsupervised learning, including clustering and association [37]. Clustering aims to find instinctive groups within a set of data, such as customer grouping according to their purchasing behavior; while association aims to explore the rules which outline big chunks of data.

Table 1 shows few comparisons between supervised learning and unsupervised learning.

Table 1. Comparison Between Supervised Learning & Unsupervised Learning

	Supervised Learning	**Unsupervised Learning**
Input Data	Trained using labeled data	Trained using unlabeled data
Type of problems	- Classification - Regression	- Clustering - Association
Algorithm	- Support Vector Machines - Random Forest - Linear Regression - Naïve Bayes - Decision Trees	- K-means - Hierarchical - Association Rules
Accuracy	- High	- Low
Application	- Spam filtering - Fraud detection	- Anomaly detection - Vehicle classification

5.4. Machine Learning in Medical Field

A lot of applications have been developed with the technology of machine learning. One of its significant applications would be within the medical field [38]. Machine learning is getting more and more popular in today's world, as it is the latest trend in enhancing the healthcare system [39]. Somehow in achieving the goal of reducing the cost of healthcare, machine learning may prove to be one of the solutions [40]. According to the research, the healthcare industry uses machine learning to diagnose a particular disease, either using patients' past records as data to predict the future outcome by entering few variables' values, or by using image processing to scan, process, and determine the future outcome [41]. Machine learning allows physicians to carry out a nearly perfect diagnosis if it is applied and utilized effectively [40]. One of the main advantages of using machine learning to predict the outcome of a particular disease is that recommendation can be made according to someone's health conditions such as his or her age, gender, and so forth, and this may assist someone to have much more health awareness on him or herself at the first stage [42]. Thus, this advantage is one of the reasons encountered in solving the problem of this proposed project.

5.5. Overview of Supervised Learning

Among various types of machine learning, supervised learning will be implemented in this project as labeled data will be used to train the proposed system. Therefore, an in-depth investigation on supervised learning algorithm will be carried out. Figure 1 below displays the overall process of carrying out supervised learning.

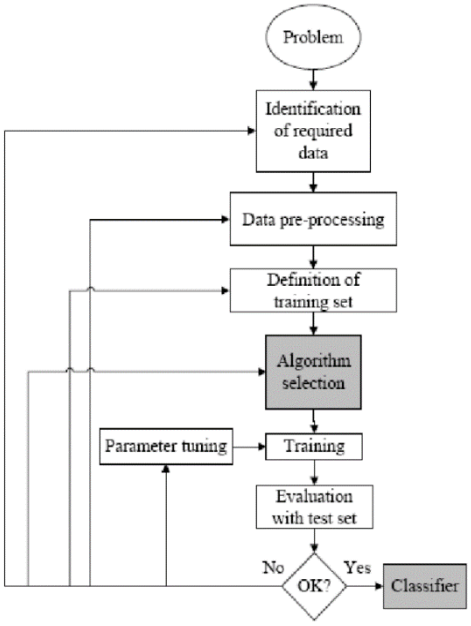

Figure 1. Overall Flowchart of Supervised Learning [43]

The first phase is the *problem identification*, which means that a specific problem has to be identified before training the algorithm [43]. For instance, figuring out whether an email is a spam email. After identifying the problem, *related data* has to be *identified* and gathered, as the data will be used to train the supervised learning algorithm [44].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Data collection is said to be one of the most significant procedures as the data's quantity as well as quality will affect the performances of the model.

The third phase would be the *pre-processing* of the data gathered. In this phase, the collected data will be loaded into an appropriate place and it will then be prepared to be used for the training process [45]. From this phase, it is time for the developer to see if there are any relationships between various variables. For instance, the developer can check whether the number of output A is more than output B. If so, the result generated would seem to be biased as the model will know that output A is correct at most of the time. Therefore, the developer has to organize the data well so that the biases of the result will not happen [45]. The gathered data will be separated into two sections, one will be used in the training process, while another one will be used in the testing process to judge the performance of the trained model [46]. The reason why the separation of data is carried out is to ensure the model would not only memorize the questions, but it is able to give an accurate result given different input.

Model selection will be carried out after pre-processing the data and defining both training and testing data. Model selection is the most important process among all, as different types of models will be used for different purposes [47]. For instance, some models are suitable for image processing, while some models are suitable for data in numerical form. After selecting the most appropriate model, *data training* process will be carried out by using the selected model [43]. The model will keep improving by keep training the model until it has the ability to get a high accuracy result [48]. In the training process, some random values will be given for weights and biases, and these values will be used to predict the output [49]. If the predicted output is not achieving high accuracy, the values of weights and biases will be kept adjusted until a high accuracy of prediction is achieved.

Evaluation is carried out when the training process is done to see whether the trained model is good enough by inputting a set of test data (which is one of the two parts separated out in the previous process) [50]. This process allows the developer to see the performance level of the trained model. If the predicted output is far different from the real output, then the developer will need to re-train the model. In contrast, if the predicted output is accurate enough, then the model is ready to be used for classification.

5.6. Comparison Between Supervised Learning Algorithms

As the proposed system will classify the output into multiple classes based on selected symptoms, therefore classification algorithm will be used. Multiple types of algorithms that can be used in solving classification problems, including support vector machines, Naïve Bayes classifier, decision trees, random forest, and so forth [51]. From the research found, a conclusion can be drawn that every developer has applied different types of supervised learning algorithms during the implementation of the prediction system.

This can be explained by the journal which aimed to compare different types of classification algorithms to detect network intrusion [52]. From the research, it can be seen that the combination of random forest and support vector machine produce the highest accuracy compared to other algorithms, such as the combination of support vector machine and BayesNet, support vector machine and logistic regression, and many more.

Another project had been done to evaluate the main reason which affects the performance of newly listed companies using support vector machine as well as various decision tree models [53]. Results showed that one of the decision tree models named C5.0 had achieved the highest accuracy with 96.46% compared to other models.

Another journal used Naïve Bayes, Decision Tree, as well as K-Nearest Neighbor algorithms to carry out sentiment classification on Roman-Urdu feedbacks [54]. From the research, it can be clearly seen that the Naïve Bayes algorithm has the best performance by achieving 97.33% accuracy for training data and 97.50% accuracy for testing data, compared to the other two algorithms.

Therefore, the researcher decided to choose three common types of multiclass classification algorithms for further investigation, including Naïve Bayes, Support Vector Machine, as well as Decision Tree classifier. Also, these three algorithms will be used to test the accuracy during the implementation, and the one with the highest accuracy will be selected to be used in the prediction process.

$$p(C_k|\mathbf{x}) = \frac{p(C_k)\ p(\mathbf{x}|C_k)}{p(\mathbf{x})} \quad (1)$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Naïve Bayes Classifier refers to a classification algorithm based on Bayes theorem with the use of probability method. It is used when the number of input variable is high [55]. Using Naïve Bayes Classifier allows the conditional probability of prediction to be calculated based on the features or input [51]. Naïve Bayes Classifier is an algorithm which assumes that every feature or variable is independent from each other. There are three different types of Naïve Bayes models, including Bernoulli, Multinomial, as well as Gaussian, and each serve with different purposes [56]. Multinomial model is used when discrete frequency count is carried out. Meanwhile, Bernoulli model is used for binary features, such as 0s and 1s. For such, the parameters are only in the form of yes or no. On the other hand, Gaussian Naïve Bayes is used when the features follow a normal distribution, such that all the features are continuous. Naïve Bayes classifier works well in various applications, including medical diagnosis, text classification, and so on [56]. Equation (1) shows the Bayes theorem's formula to calculate the probability.

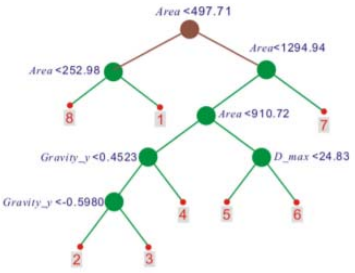

Figure 2. Example of Decision Tree Structure [57]

Decision Tree Classifier is one of the supervised machine learning algorithms which works well in classification problem [58]. It works by separating out the data continuously based on a particular parameter until a predicted outcome is generated [59]. The classifier uses the structure of tree to serve as the rules of classification by having three basic components, including nodes, branches, and leaf nodes [57]. The node in the decision tree refers to a question generated based on the features or variables available, while the leaf refers to the outcome of the question generated [58]. There are two types of decision trees, including classification tree and regression tree [60]. Classification tree is used when the predicted outcome is in categorical or discrete form. Classification tree classifier is able to solve both binary and multiclass classification problems. Meanwhile, regression tree is used when the predicted outcome is generated from continuous values. Figure 2 shows the structure of the decision tree algorithm.

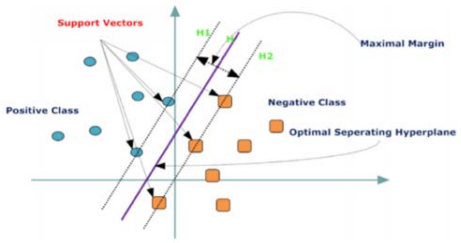

Figure 3. Example of SVM's Structure [61]

Support Vector Machine Classifier has been widely used in multiple applications [55]. Other than classification, support vector machine has been used to serve other purposes, such as ranking and regression. In classification, it categorizes multiple classes by generating a separating hyperplane [62]. In other words, support vector machine will generate the best hyperplane which separates out a chunk of data. To reduce the chance of error, margin should be maximized so that the distance between separated classes and the hyperplane is as far as possible [55]. Figure 3 shows the overall structure of linear support vector machine algorithm.

Table 2. Comparison Between 3 Machine Learning Models [55], [59], [63]

	Naïve Bayes	Decision Tree	Support Vector Machine
Advantage (s)	Simple prediction Well-performed in multiclass classification Less computational time is needed for the training process	Simple and quick Eliminate insignificant variables Interpretation is easy	Works efficiently if the separation between multiple classes are clear Memory efficient Works well in high dimensional spaces
Disadvantage (s)	Require large dataset to get better prediction result	Chance of getting overfit is high Time consuming	Time consuming Needs of huge matrix operations

5.7. Similar Systems

The study aimed to predict diabetes disease using data mining technique [64]. The system created will

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

generate an intelligent therapeutic option emotionally supportive network as an assistant for the physicians. After collecting diabetes patient's information, this information will be sent for training, testing, and predicting. Multiple algorithms are used and compared in terms of their accuracy, including Bayesian as well as K-Nearest Neighbor (KNN). The target user of this system is the administrator, and he or she will use the system to predict whether a patient has diabetes by asking the current records from patients. The administrator is allowed to choose the type of machine learning algorithm for prediction. After processing, the predicted result will be generated, and the report can be printed out and given to the patient upon request. Suggestions will be given if the patient is being diagnosed with diabetes. Therefore, before using this system, the clinical records such as serum ins, tri-fold trick, pg concentration, and so forth should be available. However, this study does not produce a real system. Instead, the study is just a concept by explaining how the proposed system will look like. Therefore, there is no accuracy result being stated since both of the algorithms are not tested yet.

A system was proposed in the year of 2015, which the study aimed to suggest an intelligent system that can assist people to diagnose whether or not they have a heart disease using machine learning and data mining techniques, including Naïve Bayes, Decision Tree, as well as Neural Network algorithm [65]. Before starting with the prediction process, current information will be taken from the patients. The study further explained the three machine learning algorithms used, as well as some basic information about heart disease. However, the system is still under research process and it has not been implemented yet.

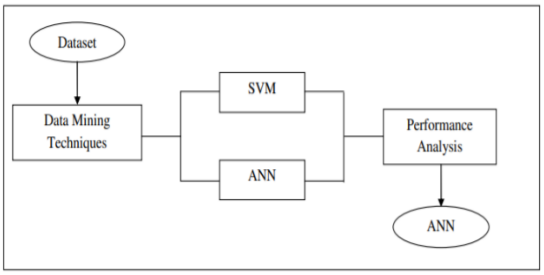

Figure 4. Flowchart of Kidney Disease Prediction [66]

In this study, prediction on kidney disease has been carried out with two machine learning algorithms, including Support Vector Machine as well as Artificial Neural Network [66]. The performance of these algorithms will be compared in terms of their accuracy and execution time. The data is gathered from medical centers, hospitals, as well as laboratories. There are 4 types of kidney diseases being used for prediction, including chronic kidney disease, chronic glomerulonephritis, acute nephritic syndrome, as well as acute renal failure. Features of the dataset include the patient's gender, age, creatinine, urea, and glomerular filtration rate. The prediction result showed that the performance of Artificial Neural Network is better than Support Vector Machine, with an accuracy of 87.70% and 76.32% respectively.

5.8. Summary

Based on the research found, it can be concluded that the traditional doctor consultation is far behind. The healthcare industry with the utilization of technologies enables all the problem contexts stated above to be solved, such as high cost, time consuming, and transportation issues. Therefore, a prediction system using machine learning and data mining technique will be able to assist patients in diagnosing disease themselves in a more simplified and convenient way. Also, the studies of similar systems prove that the proposed solutions have existed. As from the three similar systems found, a summary can be made that most of the prediction systems focus more on the data mining process instead of having a complete system with the implementation of the user interface. Therefore, it can be further strengthened by executing the proposed project.

6. METHODOLOGY

System development methodology refers to a framework that can be used in system development planning, managing, and controlling [67]. There are various types of methodologies that evolved with their strengths and weaknesses, and the selection of suitable methodology needs to look into multiple aspects of a project, such as project size, project development duration, project team size, and so forth [68]. Two different methodologies will be compared, and the best one will be used to carry out the project.

According to the research, agile digital transformation, extreme programming, and rapid application development are the three top methodologies used in software development process in the year of 2018 and they are estimated to be the best methodologies in software development

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

process in the year of 2019 [69]. In conjunction with the project scenario, both extreme programming and rapid application development are selected to be further compared. After comparing these methodologies, rapid application development (RAD) is selected to be implemented in the proposed project.

6.1. Rapid Application Development (RAD)

Generally, rapid application development is one of the agile software development methodologies which allows user's requirements to be improved or changed [70]. RAD aims to reduce the development time as well as the cost by involving users in every process [71]. The reason why RAD has been widely used in software development process is the nature of handling tight timelines and it focuses on prototyping for customers [72]. Also, as it is prototype and customer-focused, developers can produce a prototype constantly and improve it based on the feedback received from users. Besides, it is said to be suitable in the development of a new system to support a new business function of a particular firm [73]. RAD is suitable to be used when the project has a focused scope and the objectives are little yet well-defined [74]. Also, it is suitable for the project in which its decisions can be made by a small number of people [75]. Using RAD requires the project to be developed within the capabilities of technology used.

RAD is chosen as it is an iterative framework which is suitable to be used in the development of new small-to-medium scale project within a limited development duration as well as cost [76]. Besides, the chosen methodology will increase the success rate of the system development and can achieve greater user satisfaction as RAD methodology works by following an iterative life cycle (starting again) as well as evolutionary (continuous improvement) [76]. Also, it is suitable for a small project team which may include six people or even lesser [77]. Based on the evaluation made, it is believed that RAD is an ideal methodology to be used within the proposed project, as the project cost is assumed to be low, project scale and the size of the project team is small, and time is limited for the project development.

In contrast, extreme programming is inappropriate to be used in the proposed project. This is because from the research found, this type of methodology focuses more on the coding process rather than the user interface. It is said to be not suitable for this project as the target user of this project will be public, specifically the patient. Therefore, it is assumed that the structure of the user interface must be more user friendly as an organized interface allows users to use the system more comfortably.

6.2. Overview of Rapid Application Development (RAD)

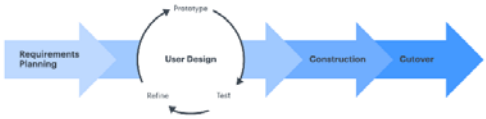

Figure 5. Process in RAD [78]

Figure 5 displays the overall process of rapid application development (RAD). From the figure above, it clearly shows that there are four phases involved in RAD, including requirement planning, user design, construction, and cutover [79].

Step 1: requirement planning

Generally, the activities carried out in requirement planning phase would be the same as the system planning and analysis phase in software development life cycle (SDLC). In this phase, the current problem faced by users will be determined. Later, the team including both users and staff agree upon the project scope as well as system requirements [80].

In relation to the proposed project, project scope (what it is supposed to be done) will be defined and all the scopes should achieve the goal of the project. When the system requirements are identified, the timeline of the project will be generated to make a plan on the task which should be done on a particular day so that the project can be done within the time given without delaying it. Other than that, cost estimation will be done as well by identifying the resources that need funds.

Step 2: user design

In RAD, the activities carried out in user design phase are similar to the system design phase in software development lifecycle (SDLC). This phase mostly emphasizing the interaction between system analyst and user [81]. A working prototype will be produced as an outcome in this phase.

In this project, the developer will start to design and produce a working prototype based on the system requirements. Users are allowed to test the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

project prototype to ensure user expectations have been met. This phase will be an iterative phase by keep collecting feedback from users so that the prototype can be improved until all users are satisfied with it. With this iteration, the developer can ensure the final product will have a higher success rate, as feedback is collected from users whenever the prototype is being updated once to meet the user's requirements. Besides, the code does not need to be changed constantly as the prototype will keep changing until it reaches high satisfaction from users before proceeding to the construction phase.

Step 3: construction

Activities carried out in RAD's construction phase would be the same as system development phase in software development lifecycle (SDLC). The development of a working system will be carried out after designing it [80]. The project team including programmers and testers will participate in this stage to ensure the system can be developed successfully by fulfilling the user's requirements. Users are allowed to participate during the process of development by suggesting improvements or changes.

In relation to the project, project implementation will be carried out in this phase, which means that the final prototype generated from the second phase will be transformed into a working model. Multiple sub-tasks are broken down, including coding and various types of testing. In this phase, the developer is assumed to complete the working system faster as majority of the changes required by users have been done at the user design phase. After completing the system, testing such as unit test and usability test will be carried out to ensure the system works well without any bugs or errors before deploying it. The system will be improved iteratively until all the errors or bugs are cleared. However, the system will still get some minor changes if users provide suggestions, new ideas, or changes after testing out the system. Therefore, the construction process will continue until it meets the user's requirements.

Step 4: cutover

The last stage of RAD would be the cutover phase, in which the activities carried out is similar to the final task in SDLC. For such, the process of testing, the changeover from the current system to a new system, as well as user training are all involved in this phase [80].

In this project, the finalized product which satisfies the user requirements will be released and deployed to the users. Generally, the company which has a current system will undergo a cutover process, by transferring the data from the current system to a new system. Since this project is a new system, there is no need to undergo the data transfer process. Whenever there are any bugs found, the system will be improved.

7. RESULTS AND DISCUSSION

7.1. Machine Learning Model

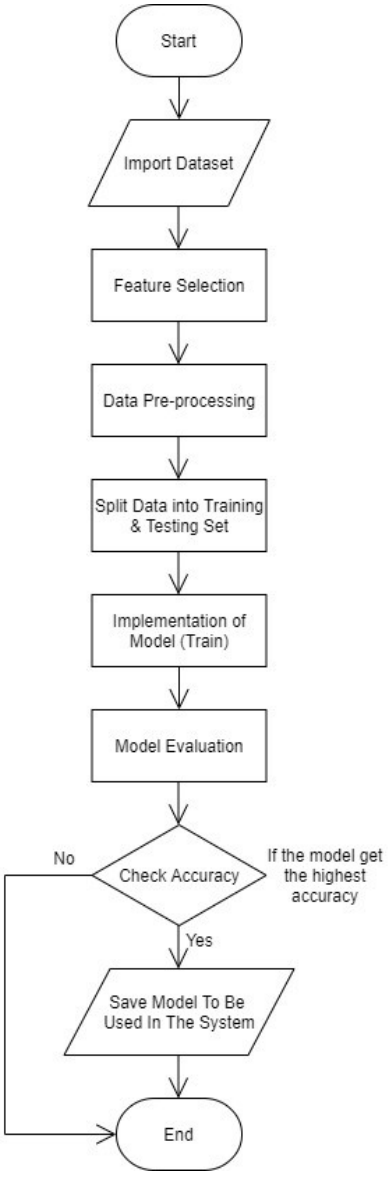

Figure 6. Flowchart of Machine Learning Model Implementation

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Figure 6 visualizes the step-by-step procedure of implementing the machine learning model within the system. Initially, the identified dataset is imported and will be used for training and testing the model. Later, feature selection will be carried out by choosing the important features which will be helpful in achieving the objective of the project, including both the dependent and independent variable. Later, pre-processing process such as the checking of noisy and missing data will be carried out, and further processes will be made if there is any. Then, the dataset will be divided into both training and testing set, and this will be done by the system randomly to avoid any biases. Three models chosen during the literature survey will be implemented, including Decision Tree, Naïve Bayes, as well as Support Vector Machine classifier. Evaluation will be done after testing out the trained model to see which model performs the best in terms of accuracy, and the most accurate model will be saved and used in the project system for further prediction.

Step 1: problem identification

As mentioned in the literature review section, there are few steps in implementing machine learning model within the system. First, to define problem of the project. The problem has been identified earlier, which aim to help the patient in improving their convenience by making disease prediction themselves anytime at anywhere.

Step 2: identification of required data

In step 2, the required data which will be helpful in achieving the project objective is identified. Since the project aimed to help patients in predicting possible disease themselves, thus the dataset used would be the most suitable one to carry out machine learning model implementation. The dataset is named as disease prediction dataset, consisting of 131 symptoms and 41 types of diseases, and it is retrieved by the author from one of the hospitals. The dataset is a multivariate dataset consisting of structured data. It is an open-source dataset that can be used by anyone. The source of the dataset is from the Kaggle website, which is one of the most reliable sources for the data scientist to get tools and resources to achieve the research goal.

Step 3: feature selection

Figure 7. Part of The Independent Variables in Dataset

In this step, the important features will be selected in implementing machine learning model, meaning the variable which is helpful in solving the project's problem will only be selected, while the useless variable will be eliminated. Figure 7 above shows part of the independent variables in the dataset. Each independent represents one symptom, therefore all attributes in the dataset is important, as it will be used in predicting the disease. In short, all the attributes in the dataset will be used as the outcome will need to depend on these independent variables heavily.

Step 4: data pre-processing

Figure 8. Summary of Attribute "Itching"

In this step, the data will be pre-processed into a proper format so that it can be understood by the machine. These steps include cleaning noisy data, missing data, as well as transforming the data into a proper data format if there are any attributes is in an improper format.

Figure 8 shows the summary of one of the attributes named "itching". From the chart displayed, it shows that the column only contains binary value, which is 0 and 1, same goes to the other attributes. Since there is no noisy data in the dataset, thus there is no need to do any further processing to clean the noisy data. Also, it can be seen that there is no missing value for each and every "cell" in the dataset. Thus, there is no need to do any processing to clean missing data.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

prognosis	prognosis
Fungal infe	1
Fungal infe	1
Fungal infe	1
Fungal infe	1
Fungal infe	1
Fungal infe	1
Fungal infe	1
Fungal infe	1
Fungal infe	1
Fungal infe	1
Allergy	2

Figure 9. Before-and-after Transformed Data

```
machinelearning_model
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 410 entries, 0 to 409
Columns: 133 entries, itching to prognosis
dtypes: int64(132), object(1)
memory usage: 424.5+ KB
None
```

Figure 10. Data Type Before Transforming

```
machinelearning_model
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 410 entries, 0 to 409
Columns: 132 entries, itching to prognosis
dtypes: int64(132)
memory usage: 422.9 KB
None
```

Figure 11. Data Type After Transforming

However, other than independent variables, the outcome of the dataset is not in an integer format, meaning it is in a categorical object format (as shown in figure 10). Since the python Scikit-learn library only allows the data to be in integer type to carry out machine learning classification model, therefore the column is transformed from object format to integer format manually. Figure 9 above shows the before-and-after transformed data for the outcome column, each categorical data will be replaced with an integer.

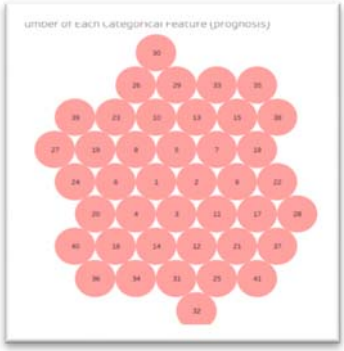

Figure 12. Visualization of Number of Data in Each Category

Also, if the number of data in each category is highly different from each other, then the data will be pre-processed. This is because an imbalance number of data will lead to the bias of data. Figure 12 above shows the number of data in each category. From the figure, it can be seen that the number of data for every category is well distributed, therefore there is no need to further pre-process on it.

Step 5: data splitting

After pre-processed data, the next step would be data splitting. In this step, all the data will be split randomly by the machine into training and testing set with the specification of data splitting ratio. The purpose of splitting data is that split data with a larger ratio will be used in training the machine learning model, while data with a lower ratio will be used in testing the machine learning model, to see whether or not the trained model performs well by looking at the accuracy. If it is not performed well, the model will be retrained.

Step 6: implementation of machine learning model

Machine learning model will be implemented in this stage. There are 3 appropriate types of multiclass classification algorithms being identified in the literature review section, including Naïve Bayes, Support Vector Machine, and Decision Tree. Therefore, these 3 machine learning algorithms will be used for training along with the pre-processed dataset together. For Naïve Bayes algorithm, there are 3 various types of algorithms, and each algorithm targets different objectives. After researching, Bernoulli Naïve Bayes is used, as this type of Naïve Bayes algorithm deals with the binary input. Since all the input of dataset chosen is in binary form, therefore Bernoulli Naïve Bayes is selected.

Step 7: model evaluation

Table 3. Accuracy of 3 Machine Learning Model

Model	Accuracy
Bernoulli Naïve Bayes	94.31%
Decision Tree	74.80%
Support Vector Machine	100%

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

In this step, the trained models (by using the 3 machine learning algorithms identified previously) are evaluated in terms of their accuracy. The model with the highest accuracy indicates that it performs the best among the three, therefore it will be implemented in the system for the disease prediction process. Table 3 summarizes the accuracy of Bernoulli Naïve Bayes, Decision Tree, and Support Vector Machine algorithm. It shows that SVM has the highest accuracy among the three, which is 100%. According to the research, support vector machine would normally perform better than the other algorithms [82]. From multiple pieces of research being done, it was found out that support vector machine provides the best by having the highest accuracy in most of the experiments [83]. Also, according to one of the researchers, support vector machine is used to carry out the classification activity on the environments by categorizing the input feature to affected and non-affected by the presence of pesticides, and the result showed that support vector machine model is able to achieve 100% accuracy. In short, support vector machine performs better than the other algorithms most of the time [84].

Step 8: save model

After evaluating the three models, it was found out that support vector machine gets the highest accuracy among the three. Thus, the model will be saved into a file called *pickle* and will be used during the disease prediction process.

7.2. Output Display from the System

Figure 13. Screenshot of the System's Main Page

Figure 13 above shows the GUI of system's main page. In this page, two buttons are being displayed so that user can choose on whether he or she wants to be loaded into the login page or signup page.

Figure 14. Screenshot of the System's Home Page

Figure 14 displays the user interface of system's home page. Users will be linked to this page after logging in successfully into one's account. In this page, guidance will be displayed so that novice user is able to use the system in a correct way.

Figure 15. Screenshot of the System's Disease Prediction Page

Figure 15 above shows the user interface of system's disease prediction page. Users will be linked to this page after clicking on the *predict* button located on the left-hand side of the menu bar. All the symptoms stored in the database will be loaded to Listbox 1, and user can add the symptoms to listbox 2 by using *add* button. *Delete* button will be used when user wants to cancel out some symptoms which have been added into listbox 2. *Predict* button will be used to carry out the prediction process by invoking the machine learning model. After prediction, user will be linked to the next page to view the predicted disease. *Cancel* button allows the user to cancel the prediction. Validation has been made in few of the buttons of the page, meaning before clicking on *add* button, the system will check whether there is a symptom being selected; if there is no symptom being selected, an error message box will be popped out to notify the user, same goes to the *delete* button. Also, added symptoms will be checked when the *add* button is clicked, to avoid redundant symptoms being added. Furthermore, the total number of symptoms will be

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

checked when the *predict* button is clicked. The validation is done here as if the number of symptoms selected by users is too less, the machine learning model may not provide an accurate diagnosis to the user.

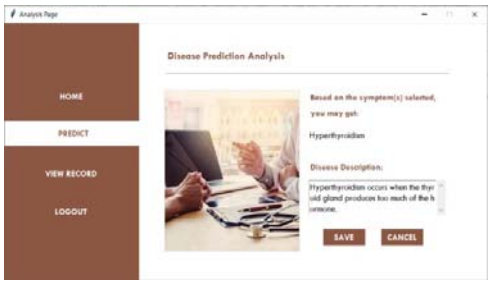

Figure 16. Screenshot of the System's Disease Prediction Result Page

Figure 16 shows the user interface of system's disease prediction result page. After pressing the *prediction* button on the previous page (disease prediction page), user will be linked to this page. The disease prediction result will be displayed to the user, together with the disease description, to let the user have a basic understanding on the predicted disease. User can choose to save the prediction record to either be viewed in the future or to be captured down and allowing the doctor to know the disease that he or she has. However, user is allowed to not save the prediction result if one wants to do so.

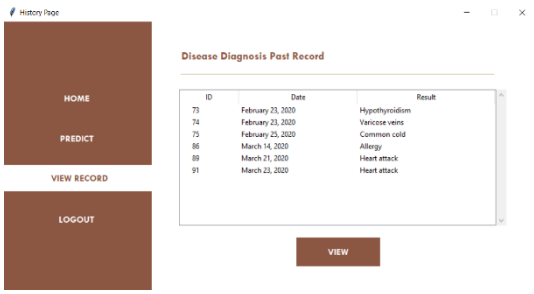

Figure 17. Screenshot of Disease Prediction History Page

Figure 17 displays the screenshot of disease prediction history page's user interface. User will be linked to this page after clicking on the *view record* button located on the left-hand side of the menu bar. In this page, user's past prediction record will be shown in summary, by displaying the diagnosis ID, diagnosis date, as well as the prediction result. *view* button allows user to view a particular record in detail when a record is selected by the user, by linking one to the next page (specific disease prediction history page).

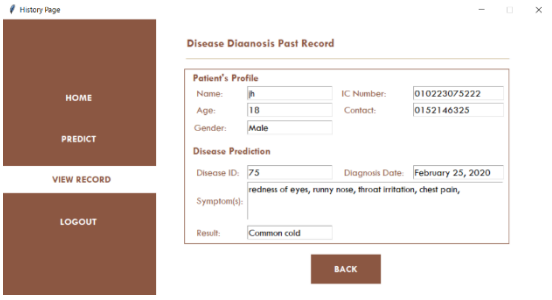

Figure 18. Screenshot of Specific Disease Prediction History Page

Figure 18 displays the user interface of specific disease prediction history page. After selecting a particular record and clicking on *view* button on the previous page (disease prediction history page), user will be linked to this page by matching the diagnosis ID of the selected record. Through this page, user can view the whole diagnosis information in detail such as user's profile information, symptoms, prediction result, and many more. User can click on the *back* button to load back to the previous page.

7.3. Summary

As mentioned at the beginning of the paper, the adoption of new technology in healthcare industry is in a rising mode nowadays, and machine learning is one of these technologies. Thus, the research contribution of this paper is to propose a new prediction system with the implementation of machine learning by allowing the patient to predict the disease themselves easily based on their symptoms. Through the implemented system, it can be explained that the adoption of machine learning in the healthcare industry is able to automate the manual doctor consultation process and eventually solve the issues mentioned previously. Other than that, since there is lack of a complete system being found throughout the research process, therefore it is believed that the proposed system can act as a testing tool for the clinical industry to diagnose the patient's disease as the researcher has displayed a more complete disease prediction system for the user to try out compared to the other existing researches which only focus on the machine learning part by testing out the accuracy of machine learning model instead of implementing a complete user interface and allowing the user to try it out by themselves.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

8. CONCLUSION AND FUTURE WORKS

The proposed system is successfully completed by reaching all the project objectives. Investigation of similar systems has been carried out to find out the problems of these similar systems. Throughout the investigation, one of the problems found is that there is lack of a complete system implemented with the data mining techniques together; meaning there are only data mining techniques being trained and tested with its accuracy without the implementation of a real system, which is not able to be used by the users. Thus, the proposed system is here to enhance the problem with an implementation of a front-end and back-end system allowing users to use by having a proper user interface.

As mentioned above, supervised learning has been used for the prediction part of this project. After training and testing out the three chosen machine learning algorithms, it can be known that support vector machine algorithm performs the best by giving 100% accuracy, while decision tree algorithm performs the worst among the three, by giving 74.80% accuracy. Thus, support vector machine is being selected and used for the disease prediction process.

Overall by having the system implemented, the research problems mentioned above can be solved gradually. The spending on manual doctor consultation can be saved as patients can use the system to predict disease and to see whether it is necessary to consult the doctor manually at a physical clinic or hospital. Also, the implementation of machine learning within the system can help the users to know the predicted disease within a short period of time and thus the rural residents can take advantage on the system as they do not need to spend time and effort in getting transportation service just to access to medication service which may be located far from home. At the same time, the system can help to increase health awareness for those who are busy with their work and have no time to consult doctor, as patients can use the system anytime to check and aware of their health condition even if they are busy.

As a whole, the system increases the chance of accessing healthcare service by patients as the system is able to give them a lot of conveniences which the issues brought up in the introduction section can all be solved. Moreover, patients do not need to have any experiences before starting to use the system as this system is designed for those who fall under novice users.

It is expected that the system being implemented can provide a significant contribution to the healthcare industry as a more complete functional prediction system has been implemented compared to the existing one which only focuses on the testing of machine learning model, and thus provides the healthcare industry as one of the options in testing the feasibility of the system in the industry.

For future enhancement, few functionalities can be added or improved. From the tester's feedback, the system can be improved by transforming the system from desktop application to mobile application. Also, more datasets will be retrieved by spending much more time in dataset searching process, allowing the researcher to have more options during the selection of the dataset.

REFERENCES

[1] Charlton, R., 2018. The importance of the 'doctor–patient relationship'. In: *Learning to Consult.* s.l.:CRC Press, pp. 44-56.

[2] Obermeyer, Z. & Emanuel, E. J., 2016. Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. *The New England journal of medicine,* 375(13), pp. 1216-1219.

[3] Deo, R. C., 2015. Machine Learning in Medicine. *Circulation,* 132(20), pp. 1920-1930.

[4] Hamet, P. & Tremblay, J., 2017. Artificial intelligence in medicine. *Metabolism,* Volume 69, pp. 36-40.

[5] Centers for Medicare and Medicaid Services, 2019. *Historical.* [Online] Available at: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NationalHealthAccountsHistorical.

[6] Deloitte, 2019. *2019 Global Health Care Outlook.* [Online] Available at: https://www2.deloitte.com/global/en/pages/life-sciences-and-healthcare/articles/global-health-care-sector-outlook.html.

[7] Slaunwhite, A. K., 2015. The Role of Gender and Income in Predicting Barriers to Mental Health Care in Canada. *Community Mental Health Journal,* 51(5), pp. 621-627.

[8] Peltzer, K. et al., 2016. Comparison of health risk behavior, awareness, and health benefit beliefs of health science and non-health science

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

students: An international study. *Nursing & Health Sciences,* 18(2), pp. 180-187.

[9] Rodriguez-Cayro, K., 2017. *How Does A Bad Job Affect You? 5 Ways Having A Bad Job Affects Your Physical & Mental Health.* [Online] Available at: https://www.bustle.com/p/how-does-a-bad-job-affect-you-5-ways-having-a-bad-job-affects-your-physical-mental-health-2941898.

[10] Atuoye, K. N. et al., 2015. Can she make it? Transportation barriers to accessing maternal and child health care services in rural Ghana. *BMC Health Services Research,* pp. 1-10.

[11] Siedlecki, R., Bem, A., Ucieklak-Jez, P. & Predkiewicz, P., 2016. Rural versus Urban Hospitals in Poland. Hospital's Financial Health. *Procedia - Social and Behavioral Sciences,* pp. 444-451.

[12] Hung, P., Henning-Smith, C. E., Casey, M. M. & Kozhimannil, K. B., 2017. Access To Obstetric Services In Rural Counties Still Declining, With 9 Percent Losing Services. *Market Concentration,* 36(9).

[13] Cordasco, K. M., Mengeling, M. A., Yano, E. M. & Washington, D. L., 2016. Health and Health Care Access of Rural Women Veterans: Findings From the National Survey of Women Veterans. *The Journal of Rural Health,* 32(4), pp. 397 - 406.

[14] Glascock, J. et al., 2018. Treatment algorithm for infants diagnosed with spinal muscular atrophy through newborn screening. *Journal of neuromuscular diseases,* 5(2), pp. 145-158.

[15] Shanks, N. H. & Buchbinder, S. B., 2016. *Introduction to Health Care Management.* 3 ed. s.l.:Jones & Bartlett Publishers.

[16] McKercher, B. & Wongkit, M., 2016. Desired Attributes of Medical Treatment and Medical Service Providers: A Case Study of Medical Tourism in Thailand. *Journal of Travel & Tourism Marketing,* 33(1), pp. 14-27.

[17] The Business Research Company, 2019. *Healthcare Global Market Opportunities And Strategies To 2022,* s.l.: Research and Markets.

[18] Taber, J. M., Leyva, B. & Persoskie, A., 2015. Why do People Avoid Medical Care? A Qualitative Study Using National Data. *Journal of general internal medicine,* 30(3), pp. 290-297.

[19] Devine, K. & O'Clock, P., 2015. An Analysis of the Benefits of Technology Implementation in the Healthcare Industry. *Journal of Health Care Finance,* 41(3).

[20] MarketWatch, 2019. *Smart Healthcare Market Size 2019, Global Trends, Industry Share, Growth Drivers, Business Opportunities and Demand Forecast to 2022.* [Online] Available at: https://www.marketwatch.com/press-release/smart-healthcare-market-size-2019-global-trends-industry-share-growth-drivers-business-opportunities-and-demand-forecast-to-2022-2019-08-26.

[21] Pramanik, M. I., Lau, R. Y., Demirkan, H. & Azad, M. A. K., 2017. Smart health: Big data enabled health paradigm within smart cities. *Expert Systems with Applications,* Volume 87, pp. 370-383.

[22] Kharel, J., Reda, H. T. & Shin, S. Y., 2017. An architecture for smart health monitoring system based. *Journal of Communications,* 12(4), pp. 228-233.

[23] Ghasemi, F., Rezaee, A. & Rahmani, A. M., 2019. Structural and behavioral reference model for IoT-based elderly health-care systems in smart home. *International Journal of Communication Systems,* 32(12), p. e4002.

[24] Ahmad, B., Khairatul, K. & Farnaza, A., 2017. An assessment of patient waiting and consultation time in a primary healthcare clinic. *An assessment of patient waiting and consultation time in a primary healthcare clinic,* 12(1), pp. 14-21.

[25] Pronsawatchai, P., Auefuea, S., Nartthanarung, A. & Soontornpipit, P., 2018. *Design of the Electronic Consultation System: Rama Health Electronic Consulting.* Krabi, IEEE, pp. 1-4.

[26] Jordan, M. I. & Mitchell, T. M., 2015. Machine learning: Trends, perspectives, and prospects. *Science,* 349(6245), pp. 255-260.

[27] Libbrecht, M. W. & Noble, W. S., 2015. Machine learning applications in genetics and genomics. *Nature Reviews Genetics 16,* pp. 321-332.

[28] Al-Jarrah, O. Y. et al., 2015. Efficient Machine Learning for Big Data: A Review. *Big Data Research,* 2(3), pp. 87-93.

[29] Brownlow, J., Zaki, M., Neely, A. & Urmetzer, F., 2015. Data and Analytics - Data-Driven Business Models: A. *The Competitive Advantage of the New Big Data World,* pp. 1-15.

[30] Blikstein, P. et al., 2014. Programming Pluralism: Using Learning Analytics to Detect Patterns in the Learning of Computer Programming. *Journal of Learning Sciences,* 23(4), pp. 561-599.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[31] Modha, D. S., 2014. *Unsupervised, supervised, and reinforced learning via spiking computation.* United States, Patent No. US8874498B2.

[32] Mueller, J. P. & Massaron, L., 2016. *Machine Leaning for Dummies.* New York: John Wiley & Sons Inc..

[33] Hardt, M., Price, E. & Srebro, N., 2016. *Equality of opportunity in supervised learning.* s.l., Neural Information Processing Systems Foundation, Inc..

[34] Junior, A. H. d. S. et al., 2015. Minimal Learning Machine: A novel supervised distance-based approach for regression and classification. *Neurocomputing,* Volume 164, pp. 34-44.

[35] MathWorks, 2019. *Unsupervised Learning.* [Online] Available at: https://www.mathworks.com/discovery/unsupervised-learning.html.

[36] Bautista, M. A., Sanakoyeu, A., Tikhoncheva, E. & Ommer, B., 2016. *CliqueCNN: Deep Unsupervised Exemplar Learning.* s.l., Neural Information Processing Systems Foundation, Inc..

[37] Celebi, M. E. & Aydin, K., 2016. *Unsupervised Learning Algorithms.* Berlin: Springer International Publishing.

[38] Asri, H., Mousannif, H., Moatassime, H. A. & Noel, T., 2016. Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. *Procedia Computer Science,* Volume 83, pp. 1064-1069.

[39] Char, D. S., Shah, N. H. & Magnus, D., 2018. Implementing Machine Learning in Health Care — Addressing Ethical Challenges. *The New England Journal of Medicine,* 378(11), pp. 981-983.

[40] Bhardwaj, R., Nambiar, A. R. & Dutta, D., 2017. A study of machine learning in healthcare. *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC),* Volume 2, pp. 236-241.

[41] Greenspan, H., Ginneken, B. v. & Summers, R. M., 2016. Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique. *IEEE Transactions on Medical Imaging,* 35(5), pp. 1153-1159.

[42] Linguraru, M. G. et al., 2016. *Device and method for classifying a condition based on image analysis.* United States, Patent No. US9443132B2.

[43] Akinsola, J. E. T., 2017. Supervised Machine Learning Algorithms: Classification and Comparison. *International Journal of Computer Trends and Technology (IJCTT),* 48(3), pp. 128-138.

[44] Hamm, J., Cao, P. & Belkin, M., 2016. Learning Privately from Multiparty Data. *International Conference on Machine Learning,* pp. 555-563.

[45] Zhao, J. & Gui, X., 2017. Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis. *IEEE Access,* Volume 5, pp. 2870-2879.

[46] Schlichtkrull, M. et al., 2018. Modeling Relational Data with Graph Convolutional Networks. *European Semantic Web Conference,* pp. 593-607.

[47] Mahdavinejad, M. S. et al., 2018. Machine learning for internet of things data analysis: a survey. *Digital Communications and Networks,* 4(3), pp. 161-175.

[48] Tanha, J., Someren, M. v. & Afsarmanesh, H., 2017. Semi-supervised self-training for decision tree classifiers. *International Journal of Machine Learning and Cybernetics,* 8(1), pp. 355-370.

[49] Ashfaq, R. A. R. et al., 2017. Fuzziness based semi-supervised learning approach for intrusion detection system. *Information Sciences,* Volume 378, pp. 484-497.

[50] Gidaris, S., Singh, P. & Komodakis, N., 2018. *Unsupervised Representation Learning by Predicting Image Rotations.* Paris, arXiv:1803.07728v1.

[51] Nayak, A. & Natarajan, S., 2016. Comparative study of Naïve Bayes, Support Vector Machine and Random Forest Classifiers in Sentiment Analysis of Twitter feeds. *International Journal of Advanced Studies in Computer Science and Engineering,* 5(1), pp. 14-17.

[52] Chand, N. et al., 2016. A comparative analysis of SVM and its stacking with other classification algorithm for intrusion detection. *2016 International Conference on Advances in Computing, Communication, & Automation (ICACCA)(Spring),* pp. 1-6.

[53] Basti, E., Kuzey, C. & Delen, D., 2015. Analyzing initial public offerings' short-term performance using decision trees and SVMs. *Decision Support Systems,* Volume 73, pp. 15-27.

[54] Bilal, M., Israr, H., Shahid, M. & Khan, A., 2016. Sentiment classification of Roman-Urdu opinions using Naïve Bayesian, Decision Tree and KNN classification techniques. *Journal of*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

King Saud University - Computer and Information Sciences, 28(3), pp. 330-344.

[55] Nikam, S. S., 2015. A Comparative Study of Classification. *Oriental Journal of Computer Science & Technology,* 8(1), pp. 13-19.

[56] Xu, S., 2018. Bayesian Naïve Bayes classifiers to text classification. *Journal of Information Science,* 44(1), pp. 48-59.

[57] Zhang, H. et al., 2015. A new method for nondestructive quality evaluation of the resistance spot welding based on the radar chart method and the decision tree classifier. *The International Journal of Advanced Manufacturing Technology,* 78(5-8), pp. 841-851.

[58] Dai, Q.-y., Zhang, C.-p. & Wu, H., 2016. Research of Decision Tree Classification Algorithm in Data Mining. *International Journal of Database Theory and Application,* 9(5), pp. 1-8.

[59] Jadhav, S. D. & Channe, H. P., 2016. Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques. *International Journal of Science and Research (IJSR),* 5(1), pp. 1842-1845.

[60] Youssef, A. M., Pourghasemi, H. R., Al-Katheeri, M. M. & Pourtaghi, Z. S., 2016. Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region, Saudi Arabia. *Landslides,* 13(5), pp. 839-856.

[61] Shen, L. et al., 2016. Evolving support vector machines using fruit fly optimization for medical data classification. *Knowledge-Based Systems,* Volume 96, pp. 61-75.

[62] Bledsoe, J. C. et al., 2016. Diagnostic classification of ADHD versus control: support vector machine classification using brief neuropsychological assessment. *Journal of attention disorders.*

[63] Goeschel, K., 2016. Reducing false positives in intrusion detection systems using data-mining techniques utilizing support vector machines, decision trees, and naive Bayes for off-line analysis. *SoutheastCon 2016,* pp. 1-6.

[64] Shetty, D., Rit, K., Shaikh, S. & Patil, N., 2017. Diabetes disease prediction using data mining. *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS),* pp. 1-5.

[65] Gandhi, M. & Singh, S. N., 2015. Predictions in heart disease using techniques of data mining. *2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE),* pp. 520-525.

[66] Vijayarani, S. & Dhayanand, S., 2015. Kidney Disease Prediction Using SVM and ANN Algorithms. *International Journal of Computing and Business Research (IJCBR),* 6(2).

[67] Matharu, G. S., 2015. Empirical study of agile software development methodologies: A comparative analysis. *ACM SIGSOFT Software Engineering Notes,* 40(1), pp. 1-6.

[68] Ahimbisibwe, A., Cavana, Y. R. & Daellenbach, U., 2015. A contingency fit model of critical success factors for software development projects. *Journal of Enterprise Information Management,* 28(1), pp. 7-33.

[69] Urias, E., 2019. *Best Software Development Methodologies In 2019.* [Online] Available at: https://invidgroup.com/best-software-development-methodologies-in-2019/.

[70] Stoica, M., Ghilic-Micu, B., Mircea, M. & Uscatu, C., 2016. Analyzing Agile Development – from Waterfall Style to Scrumban. *Informatica Economică,* 20(4), pp. 5-14.

[71] Abdulwahab, L. et al., 2015. *The Third International Conference on Digital Enterprise and Information Systems (DEIS2015).* Shenzhen, The Society of Digital Information and Wireless Communications (SDIWC).

[72] Soni, D. & Kohli, P., 2017. Cost Estimation Model for Web Applications using Agile Software. *Science & Technology,* 25(3), pp. 931-938.

[73] Golovin, D., 2017. OutSystems as a Rapid. pp. 1-40.

[74] ProjectManagement.com, 2019. *Process/Project RAD - RAD - Rapid Application Development Process.* [Online] Available at: https://www.projectmanagement.com/content/processes/11306.cfm.

[75] Agrawal, S., 2019. Using Rapid Application Development For Software Development Projects. pp. 1-97.

[76] Hassan, S., Qamar, U. & Idris, M. A., 2015. *Purification of requirement engineering model for rapid application development.* Beijing, China, IEEE, pp. 357-362.

[77] Cronin, B. et al., 2017. *Enabling Rapid Integration of Combined Arms Teams into a Brigade Combat Team Organizational Structure,* United States: ICF

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

INTERNATIONAL INC FAIRFAX VA FAIRFAX.

[78] Lucidchart Content Team, 2018. *4 Phases of Rapid Application Development Methodology.* [Online] Available at: https://www.lucidchart.com/blog/rapid-application-development-methodology.

[79] Shaydulin, R. & Sybrandt, J., 2017. To Agile, or not to Agile: A Comparison of Software Development Methodologies. pp. 1-11.

[80] Maheshwaran, P., kumar, R., Rajeswari, S. & Mungara, J., 2017. A Review On Requirement Engineering in Rapid Application. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 2(3), pp. 742-746.

[81] Ali, K., 2017. Ali, Kazim. "A Study of Software Development Life Cycle Process Models. *International Journal of Advanced Research in Computer Science,* 8(1).

[82] Peter, S. C. et al., 2019. Quantitative Structure-Activity Relationship (QSAR): Modeling Approaches to Biological Applications. *Encyclopedia of Bioinformatics and Computational Biology,* Volume 2, pp. 661-676.

[83] Noi, P. T., 2018. Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery. *Sensors,* 18(1), p. 18.

[84] Niell, S. et al., 2018. Beehives biomonitor pesticides in agroecosystems: Simple chemical and biological indicators evaluation using Support Vector Machines (SVM). *Ecological Indicator,* Volume 91, pp. 149-154.