**WEB VULNERABILITY DETECTION USING MACHINE LEARNING**

**CHAPTER ONE**

**1.0 Introduction**

**1.1 Background Of Study**

The security of the web application is most important, according to the report by Web Application Security Consortium, about 49% web application contains the high severity level vulnerabilities and 13% of them are automatically get compromised for the security vulnerabilities. The unsecure web application leads to the known security vulnerabilities such as SQL Injection, Cross-side scripting, cookie theft, security misconfiguration, session hijacking, self-propagating worm's attacks, etc (Sonali. 2019)

SQL Injection can also be used to add, modify and delete records in a database, affecting data integrity. To such an extent, SQL Injection can provide an attacker with unauthorized access to sensitive data. SQL injection is a code injection technique, used to attack data-driven applications, in which malicious SQL statements are inserted into an entry field for execution (e.g. to dump the database contents to the attacker). SQL injection must exploit a security vulnerability in an application's software, for example, when user input is either incorrectly filtered for string literal escape characters embedded in SQL statements or user input is not strongly typed and unexpectedly executed.

An SQL injection is a computer attack in which malicious code is embedded in a poorly-designed application and then passed to the backend database. The malicious data then produces database query results or actions that should never have been executed. A SQL injection (SQLi) is a type of security exploit in which the attacker adds Structured Query Language (SQL) code to a Web form input box in order to gain access to unauthorized resources or make changes to sensitive data. (Suhaimi, 2011)

Machine learning is a branch of artificial intelligence and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy. Machine learning is an important component of the growing field of data science. Through the use of statistical methods, algorithms are trained to make classifications or predictions, and to uncover key insights in data mining projects. These insights subsequently drive decision making within applications and businesses, ideally impacting key growth metrics. As big data continues to expand and grow, the market demand for data scientists will increase. (IBM, 2021)

**1.2 Statement Of The Problem**

Nowadays most activities are done online and, in the process, sensitive information’s are shared on these platforms, it is only important that this information is protected from the hands of unauthorized users who might seek to use it for questionable purposes.

This problem will help to detect and mitigate the gap in which these negative users will seek to exploit to gather this sort of sensitive information. This project serves a guide to predict and also detect malicious activities across several websites.

**1.3 Aim Of The Study**

The aim of the project is to develop a web vulnerability detection system using machine learning, this will play an important role in sql detection and other web vulnerabilities.

**1.4 Objective Of The Study**

* Design a model for detecting web vulnerabilities.
* Implement the model to perform the task.
* Test the system developed.

**1.5 Proposed Methodology**

Requirements for detecting web vulnerabilities in this project includes the use of ensemble learning that is combining five algorithms (Random Forest, Support Vector Machine, Decision Tree, K Nearest Neighbour, Naïve bayes algorithm). This is chosen because of its ability to analyse behaviours and modelling prediction to deal the web vulnerabilities and thereby making internet safer.

A model for this project was designed using the Unified Modelling Language. The system was then implemented using Python language. The system will then be tested by users to check the features of the application.

**1.6 Significance Of The Project**

This project is of great necessity due to the significant rise in fraud and the extent to which it affects the lives of those affected. The project proposes an efficient system to predict and properly annul these issues.

* The need to make the internet safer
* The need to protect sensitive information e.g., credit card details and so on.
* It would increase the level of trust customers have in credit cards.

**1.7 Scope Of The Study**

The scope of this project was limited to web vulnerability detection using machine learning.

**1.8 Organization Of Project**

Chapter one discusses the background of the project work, the problem statement and the methodology which were carried out. Chapter two provides the literature review on the work and some related works. Chapter three explains the system analysis and design of the work in detail. Chapter four discusses the implementation of the project. Chapter five summarizes the project and provides recommendations for further development of the project.

**CHAPTER TWO**

**LITERATURE REVIEW**

**2.1 Introduction**

This chapter contains the review of literature related to the project

**2.2.1 What is web application?**

Web application is a distributed application which enables the dynamic information and service delivery. Web application consists of client side and server side components. The client side component which includes static web pages with embedded scripting languages. Client can communicate with server side code using asynchronous call such as AJAX and dynamically updates the HTML pages. The web applications interact with the backend file system or database server for storing and retrieving data. Aspects of web application including programming language, state maintenance and logic implementation differentiate the web application from traditional applications.

**2.2.2 Web Vulnerabilities**

A **website vulnerability** is a weakness or misconfiguration in a website or web application code that allows an attacker to gain some level of control of the site, and possibly the hosting server. Most **vulnerabilities** are exploited through automated means, such as vulnerability scanners and botnets. (Site Lock, 2017)

2.3 Types of web vulnerabilities

### ****2.3.1 SQL Injections****

SQL injection is a type of web application security vulnerability in which an attacker attempts to use application code to access or corrupt database content. If successful, this allows the attacker to create, read, update, alter, or delete data stored in the back-end database. SQL injection is one of the most prevalent types of web application vulnerabilities.

### ****2.3.2 Cross Site Scripting (XSS)****

Cross-site scripting (XSS) targets an application's users by injecting code, usually a client-side script such as JavaScript, into a web application's output. The concept of XSS is to manipulate client-side scripts of a web application to execute in the manner desired by the attacker. XSS allows attackers to execute scripts in the victim's browser which can hijack user sessions, deface websites or redirect the user to malicious sites.

### ****2.3.3. Broken Authentication & Session Management****

Broken authentication and session management encompass several security issues, all of them having to do with maintaining the identity of a user. If authentication credentials and session identifiers are not protected at all times, an attacker can hijack an active session and assume the identity of a user.

### ****2.3.4. Insecure Direct Object References****

Insecure direct object reference is when a web application exposes a reference to an internal implementation object. Internal implementation objects include files, database records, directories and database keys. When an application exposes a reference to one of these objects in a URL, hackers can manipulate it to gain access to a user's personal data.

### ****2.3.5. Security Misconfiguration****

Security misconfiguration encompasses several types of vulnerabilities all centered on a lack of maintenance or a lack of attention to the web application configuration. A secure configuration must be defined and deployed for the application, frameworks, application server, web server, database server and platform. Security misconfiguration gives hackers access to private data or features and can result in a complete system compromise.

### ****2.3.6. Cross-Site Request Forgery (CSRF)****

Cross-Site Request Forgery (CSRF) is a malicious attack where a user is tricked into performing an action he or she didn't intend to do. A third-party website will send a request to a web application that a user is already authenticated against (e.g., their bank). The attacker can then access functionality via the victim's already authenticated browser. Targets include web applications like social media, in browser email clients, online banking, and web interfaces for network devices.

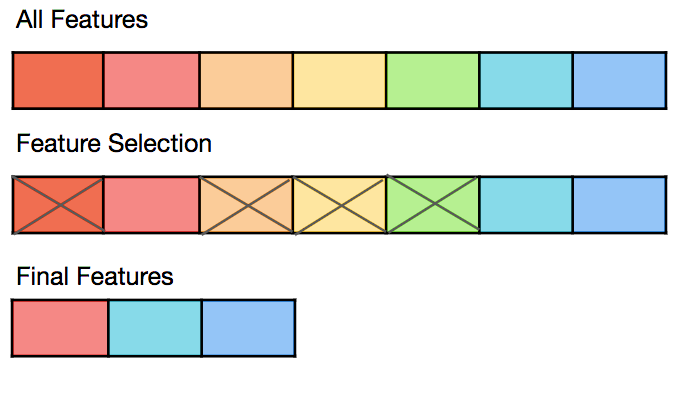
**2.4 Machine Learning**

Machine learning is a branch of artificial intelligence and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy. Machine learning is an important component of the growing field of data science. Through the use of statistical methods, algorithms are trained to make classifications or predictions, and to uncover key insights in data mining projects. These insights subsequently drive decision making within applications and businesses, ideally impacting key growth metrics. As big data continues to expand and grow, the market demand for data scientists will increase. (IBM, 2021)

**2.4.1 MACHINE LEARNING BREAKDOWN**

### 2.4.2 Feature Selection

Feature Selection is the means of selecting variables that are useful for model training and development. It helps identify which features are important in the construction of predictive models which is done to reduce the number of input variables that, in turn, reduce the modeling computational cost and, in this case, improve the model's performance.



**Figure 2.1 Feature Selection**

Categories of feature selection methods are:

1. **Feature-based feature selection:** This method involves evaluating the relationship between each input variable and the target variable using statistics and selecting those input variables that have the strongest relationship with the target variable. These methods can be fast and effective, although the choice of statistical measures depends on the data type of both the input and output variables.
2. **Filter based:** We specify some metric and based on that filter features. Correlation / chi-square could be an example of such a metric.
3. **Wrapper-based:** Wrapper methods consider the selection of a set of features as a search problem. Example: Recursive Feature Elimination
4. Embedded: Embedded methods use algorithms that have built-in feature selection methods. For example, Lasso and RF have their own method of selecting features.

Some examples of feature selection are:

1. Boruta
2. Variable Importance from Machine Learning Algorithms
3. Lasso Regression
4. Step wise Forward and Backward Selection
5. Relative Importance from Linear Regression
6. Recursive Feature Elimination (RFE)

**2.4.2.1 Feature Extraction**

Feature extraction is a dimension reduction process which reduces the initial set of raw data to more manageable groups for processing. A characteristic feature of these large data sets is many variables that require a lot of computing resources to process. Feature extraction selects and/or combines variables into features, effectively reducing the amount of data to be processed while still accurately and fully describing the original data set.

The extraction process reduces the number of resources needed for processing without losing important or relevant information. It may also reduce the amount of redundant data for a given analysis and in addition, reduce the data and the effort of the computer to create variable combinations (features).

Advantages of feature extraction are, accuracy improvements, overfitting risk reduction, speed up in training, improved Data Visualization and an increase in explainability of the model.

### 2.4.3.1 Algorithm Families

Algorithms are often grouped by similarity in terms of their function (how they work) (Jason Brownlee, 2019). It is a very useful way to group algorithms, and it is the approach that is adopted in the project.

It is not a perfect grouping method because there are still algorithms that could just as easily fit into multiple categories like Learning Vector Quantization which is both a neural network method and an instance-based method.

Some of these groupings methods are:

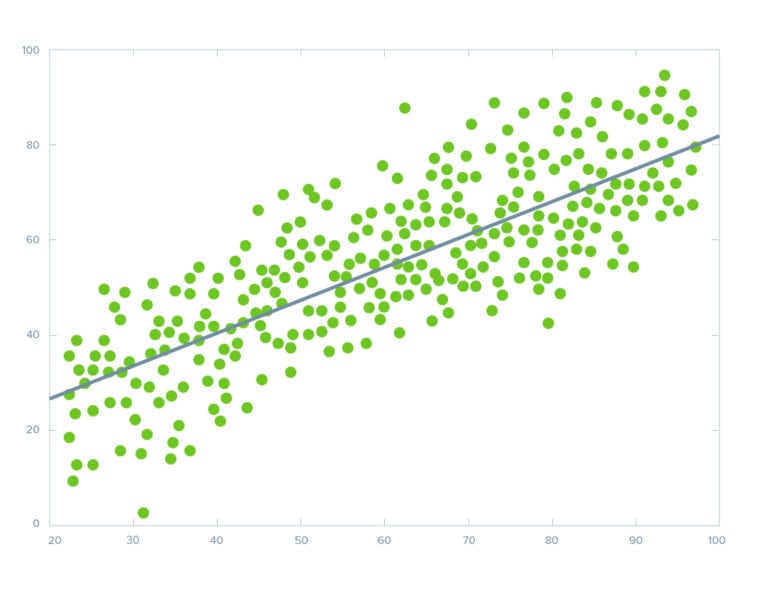
1. **Regression Algorithms**

Regression is concerned with modelling the relationship between variables that is iteratively refined using a measure of error in the predictions made by the model.

Regression methods are a workhorse of statistics and have been co-opted into statistical machine learning. This may be confusing because we can use regression to refer to the class of problem and the class of algorithm. Really, regression is a process.

The most popular regression algorithms are:

1. Ordinary Least Squares Regression (OLSR)
2. Linear Regression
3. Logistic Regression
4. Stepwise Regression
5. Multivariate Adaptive Regression Splines (MARS)
6. Locally Estimated Scatterplot Smoothing (LOESS)



**Figure 2.2 Regression Algorithms**

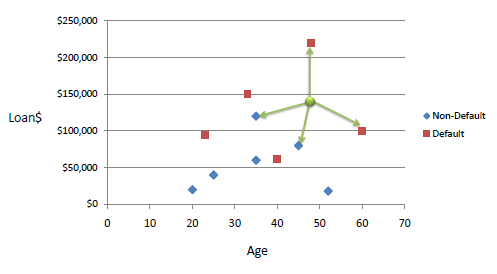
1. **Instance-based Algorithms**

Instance-based Algorithms Instance-based learning model is a decision problem with instances or examples of training data that are deemed important or required to the model.

Such methods typically build up a database of example data and compare new data to the database using a similarity measure in order to find the best match and make a prediction. For this reason, instance-based methods are also called winner-take-all methods and memory-based learning. Focus is put on the representation of the stored instances and similarity measures used between instances.

The most popular instance-based algorithms are:

1. k-Nearest Neighbour (KNN)
2. Learning Vector Quantization (LVQ)
3. Self-Organizing Map (SOM)
4. Locally Weighted Learning (LWL)
5. Support Vector Machines (SVM)



**Figure 2.3 Instance-Based Algorithms**

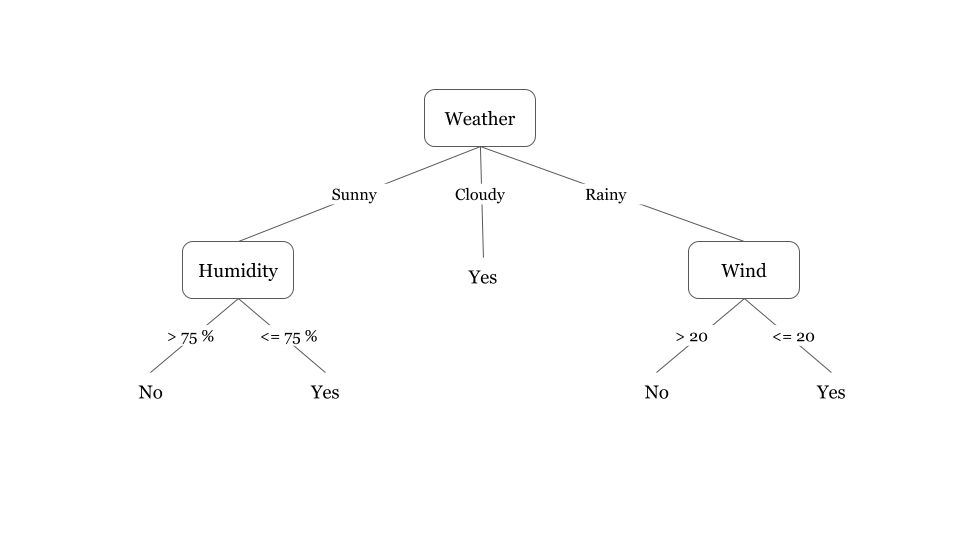
1. **Decision Tree Algorithms**

Decision tree methods construct a model of decisions made based on actual values of attributes in the data.

Decisions fork in tree structures until a prediction decision is made for a given record. Decision trees are trained on data for classification and regression problems. Decision trees are often fast and accurate and a big favorite in machine learning.

The most popular decision tree algorithms are:

1. Classification and Regression Tree (CART)
2. Iterative Dichotomiser 3 (ID3)
3. C4.5 and C5.0 (different versions of a powerful approach)
4. Chi-squared Automatic Interaction Detection(CHAID)
5. Decision Stump
6. M5
7. Conditional Decision Trees
8. Random forest



**Figure 2.4 Decision Tree Algorithms**

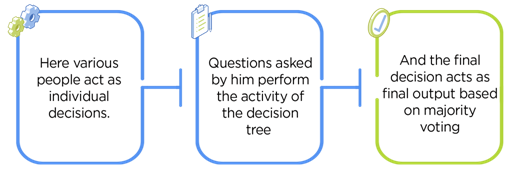
**Vii. Random Forest Explained**

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

One of the most important features of the Random Forest Algorithm is that it can handle the data set containing **continuous variables** as in the case of regression and **categorical variables** as in the case of classification. It performs better results for classification problems.

## **Real Life Analogy**

Let’s dive into a real-life analogy to understand this concept further. A student named X wants to choose a course after his 10+2, and he is confused about the choice of course based on his skill set. So he decides to consult various people like his cousins, teachers, parents, degree students, and working people. He asks them varied questions like why he should choose, job opportunities with that course, course fee, etc. Finally, after consulting various people about the course he decides to take the course suggested by most of the people. (Sruthi, 2021)

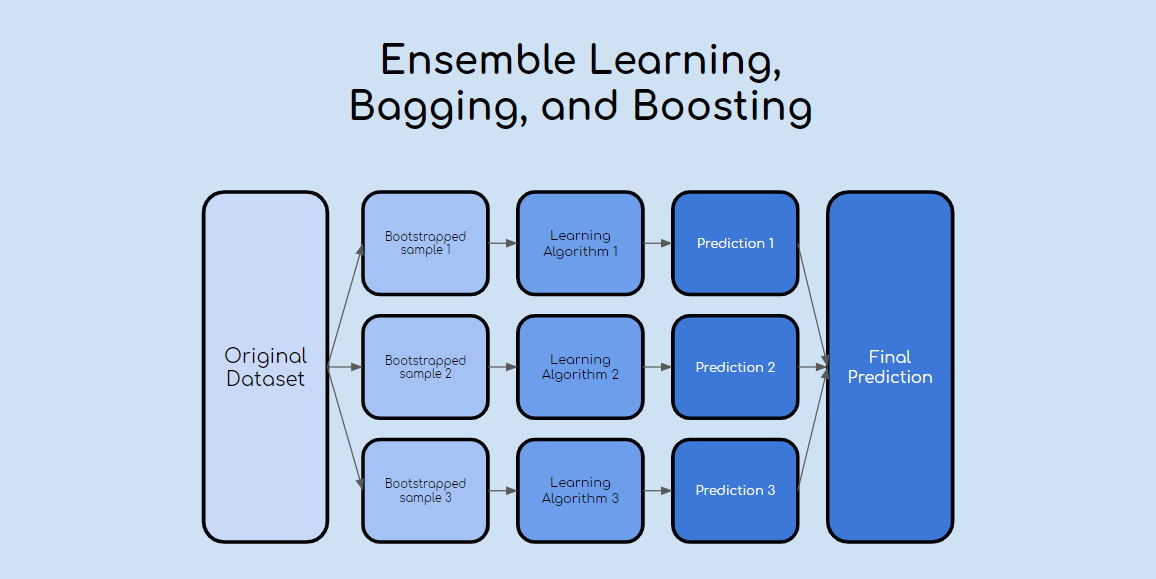


**Fig 2.5 Random Forest**

## **Viii. Ensemble Learning**

The ensemble methods in machine learning combine the insights obtained from multiple learning models to facilitate accurate and improved decisions. These methods follow the same principle as the example of buying an air-conditioner cited above. In learning models, noise, variance, and bias are the major sources of error. The ensemble methods in machine learning help minimize these error-causing factors, thereby ensuring the accuracy and stability of machine learning (ML) algorithms. Ensemblesimplymeans combining multiple models. (Simplilearn, 2021)

Thus, a collection of models is used to make predictions rather than an individual model.

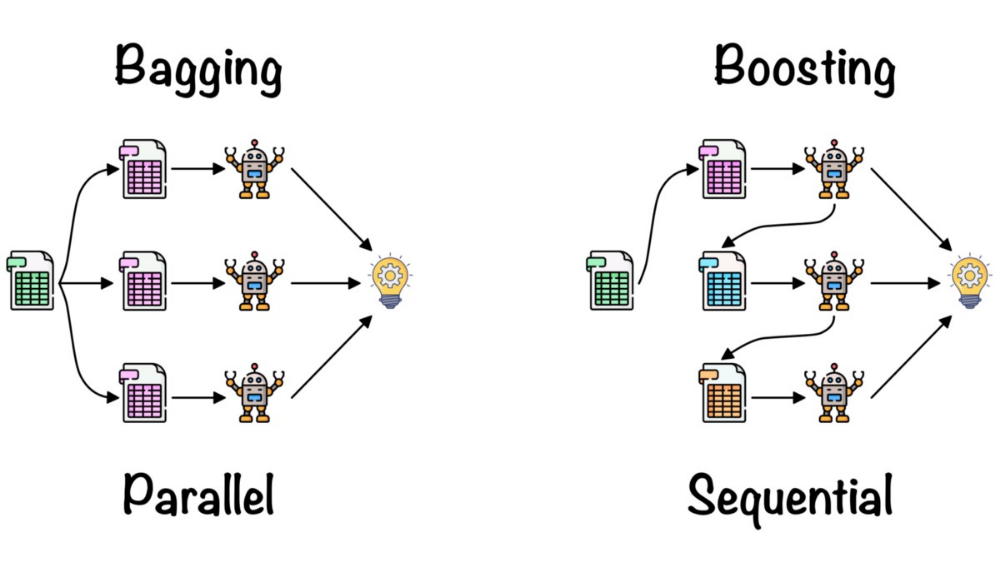


**Fig 2.6 Ensemble learning explained diagrammatically**

**Ensemble Techniques:**

1. **Bagging**– It creates a different training subset from sample training data with replacement & the final output is based on majority voting. For example, Random Forest.

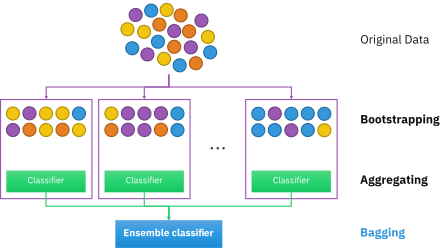
2. **Boosting**– It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. For example, ADA BOOST, XG BOOST



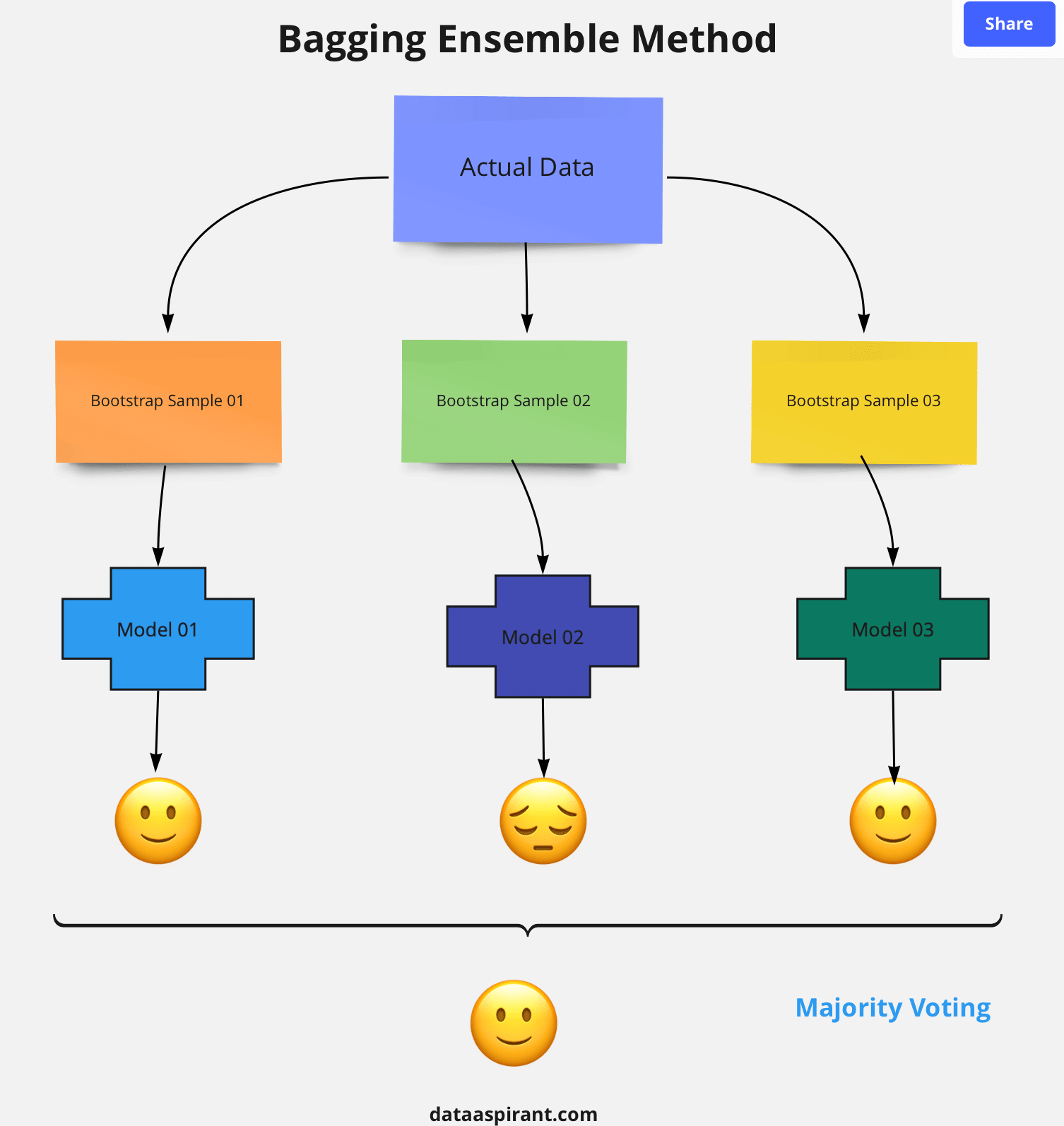
As mentioned earlier, Random forest works on the Bagging principle. Now let’s dive in and understand bagging in detail. (Agraval, 2021)

#### **Bagging**

Bagging, also known as Bootstrap Aggregation is the ensemble technique used by random forest. Bagging chooses a random sample from the data set. Hence each model is generated from the samples (Bootstrap Samples) provided by the Original Data with replacement known as row sampling. This step of row sampling with replacement is called bootstrap. Now each model is trained independently which generates results. The final output is based on majority voting after combining the results of all models. This step which involves combining all the results and generating output based on majority voting is known as aggregation.



Now let’s look at an example by breaking it down with the help of the following figure. Here the bootstrap sample is taken from actual data (Bootstrap sample 01, Bootstrap sample 02, and Bootstrap sample 03) with a replacement which means there is a high possibility that each sample won’t contain unique data. Now the model (Model 01, Model 02, and Model 03) obtained from this bootstrap sample is trained independently. Each model generates results as shown. Now Happy emoji is having a majority when compared to sad emoji. Thus based on majority voting final output is obtained as Happy emoji. (Ragav, 2017)



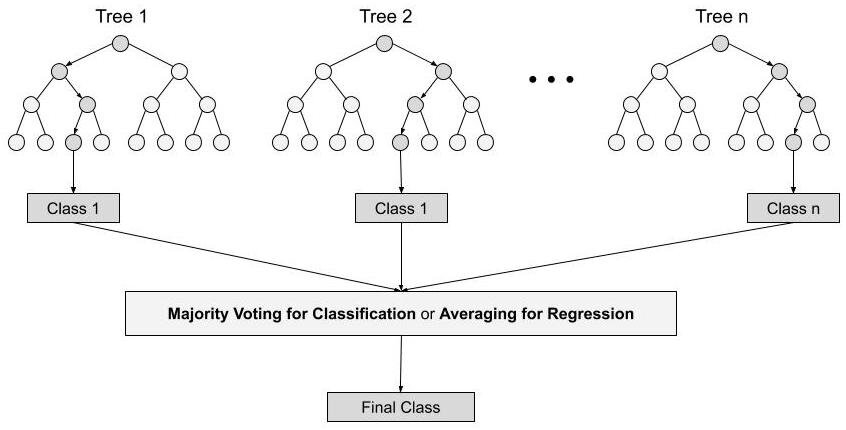
**Steps involved in random forest algorithm:**

Step 1: In Random forest n number of random records are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

Step 4: Final output is considered based on **Majority Voting or Averaging**for Classification and regression respectively. (Ellen, 2021)



For example:  consider the fruit basket as the data as shown in the figure below. Now n number of samples are taken from the fruit basket and an individual decision tree is constructed for each sample. Each decision tree will generate an output as shown in the figure. The final output is considered based on majority voting. In the below figure you can see that the majority decision tree gives output as an apple when compared to a banana, so the final output is taken as an apple.



### Important Features of Random Forest

**1. Diversity-**Not all attributes/variables/features are considered while making an individual tree, each tree is different.

**2. Immune to the curse of dimensionality-** Since each tree does not consider all the features, the feature space is reduced.

**3. Parallelization-**Each tree is created independently out of different data and attributes. This means that we can make full use of the CPU to build random forests. (Sruthi, 2021)

**4.  Train-Test split-**In a random forest we don’t have to segregate the data for train and test as there will always be 30% of the data which is not seen by the decision tree.

**5.  Stability-**Stability arises because the result is based on majority voting/ averaging.

### Difference Between Decision Tree & Random Forest

Random forest is a collection of decision trees; still, there are a lot of differences in their behavior.

|  |  |
| --- | --- |
| **Decision trees** | **Random Forest** |
| 1. Decision trees normally suffer from the problem of overfitting if it’s allowed to grow without any control. | 1. Random forests are created from subsets of data and the final output is based on average or majority ranking and hence the problem of overfitting is taken care of. |
| 2. A single decision tree is faster in computation. | 2. It is comparatively slower. |
| 3. When a data set with features is taken as input by a decision tree it will formulate some set of rules to do prediction. | 3. Random forest randomly selects observations, builds a decision tree and the average result is taken. It doesn’t use any set of formulas. |

Thus random forests are much more successful than decision trees only if the trees are diverse and acceptable. (Sruthi, 2021)

1. **Bayesian Algorithms**

Bayesian methods are those that explicitly apply Bayes’ Theorem for problems such as classification and regression.

The most popular Bayesian algorithms are:

1. Naive Bayes
2. Gaussian Naive Bayes
3. Multinomial Naive Bayes
4. Averaged One-Dependence Estimators (AODE)
5. Bayesian Belief Network (BBN)
6. Bayesian Network (BN)

A screenshot of a cell phone

Description automatically generated

**Figure 2.5 Bayesian Algorithms**

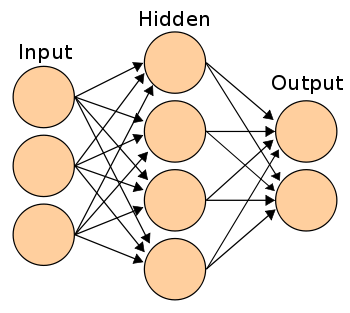
1. **Artificial Neural Network Algorithms**

Artificial Neural Networks are models that are inspired by the structure and/or function of biological neural networks.

They are a class of pattern matching that are commonly used for regression and classification problems but are really an enormous subfield comprised of hundreds of algorithms and variations for all manner of problem types.

The most popular artificial neural network algorithms are:

1. Perceptron
2. Multilayer Perceptron (MLP)
3. Back-Propagation
4. Stochastic Gradient Descent
5. Hopfield Network
6. Radial Basis Function Network (RBFN) (Amir, 2021)



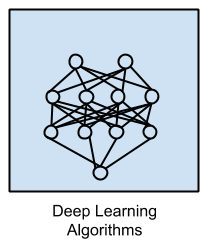
**Figure 2.6 Artificial Neural Network Algorithms**

1. **Deep Learning Algorithms**

Deep Learning methods are a modern update to Artificial Neural Networks that exploit abundant cheap computation.

They are concerned with building much larger and more complex neural networks and, as commented on above, many methods are concerned with very large datasets of labelled analog data, such as image, text. audio, and video. The most popular deep learning algorithms are:

1. Convolutional Neural Network (CNN)
2. Recurrent Neural Networks (RNNs)
3. Long Short-Term Memory Networks (LSTMs)
4. Stacked Auto-Encoders
5. Deep Boltzmann Machine (DBM)
6. Deep Belief Networks (DBN)



**Figure 2.7 Deep learning Algorithms**

Other Machine Learning Algorithms:

1. Clustering Algorithms
2. Feature selection algorithms
3. Algorithm accuracy evaluation
4. Performance measures
5. Regularization Algorithms
6. Optimization algorithms
7. Ensemble Algorithms
8. Association Rule Learning Algorithms
9. Dimensionality Reduction Algorithms

### Model Evaluation (Performance Metrics

Model evaluation is intended to estimate a model's generalization accuracy on future data (unseen / out of sample). To estimate model efficiency, performance metrics are used. Choosing evaluation criteria depends on a task of machine learning (such as grouping, regression, ranking, clustering, subject modelling, etc.).

By using different performance assessment metrics, we should be able to improve our model's overall predictive power before testing it out on unknown data for output.

Different performance metrics are used for different kinds of problems resulting in the development and selection of a model that provides high precision from sample data.

The different types of evaluation metrics available are:

1. Accuracy, Precision, and Recall
2. Confusion Matrix
3. F1 Score
4. Gain and Lift Charts
5. Kolmogorov Smirnov Chart
6. AUC – ROC
7. Log Loss
8. Gini Coefficient
9. Concordant – Discordant Ratio
10. Root Mean Squared Error
11. Cross Validation
12. Sensitivity & Specificity etc.

**2.5 Review Of Existing Works**

Mohd Yunus, Mohd Amin & Brohan, Muhammad & Mohd Nawi, Nazri & Salwana, Ely & Najib, Nurhakimah & Liang, Chan. (2018). **Review of SQL Injection : Problems and Prevention. JOIV : International Journal on Informatics Visualization. 2. 215. 10.30630/joiv.2.3-2.144**. SQL injection happened in electronic records in database and it is still exist even after two decades since it first happened. Most of the web-based applications are still vulnerable to the SQL injection attacks. Although technology had improved a lot during these past years, but, hackers still can find holes to perform the SQL injection. There are many methods for this SQL injection to be performed by the hackers and there is also plenty of prevention for the SQL injection to be happened. The vulnerability to SQL injection is very big and this is definitely a huge threat to the web based application as the hackers can easily hacked their system and obtains any data and information that they wanted anytime and anywhere. This paper can conclude that several proposed techniques from existing journal papers used for preventing SQL injection. Then, it comes out with Blockchain concept to prevent SQL injection attacks on database management system (DBMS) via IP.

Barud Demilie, Wubetu & Gizachew, Fitsum. (2022). **Detection and prevention of SQLI attacks and developing compressive framework using machine learning and hybrid techniques. Journal of Big Data. 9. 10.1186/s40537-022-00678-0.** A web application is a software system that provides an interface to its users through a web browser on any operating system (OS). Despite their growing popularity, web application security threats have become more diverse, resulting in more severe damage. Malware attacks, particularly SQLI attacks, are common in poorly designed web applications. This vulnerability has been known for more than two decades and is still a source of concern. Accordingly, different techniques have been proposed to counter SQLI attacks. However, the majority of them either fail to cover the entire scope of the problem. The structured query language injection (SQLI) attack is among the most harmful online application attacks and often happens when the attacker(s) alter (modify), remove (delete), read, and copy data from database servers. All facets of security, including confidentiality, data integrity, and data availability, can be impacted by a successful SQLI attack. This paper investigates common SQLI attack forms, mechanisms, and a method of identifying, detecting, and preventing them based on the existence of the SQL query. Here, we have developed a comprehensive framework for detecting and preventing the effectiveness of techniques that address specific issues following the essence of the SQLI attacks by using traditional Navies Bayes (NB), Decision Trees (DT), Support Vectors Machine (SVM), Random Forests (RF), Logistic Regression (LR), and Neural Networks Based on Multilayer Perceptron (MLP), and hybrid approach are used for our study. The machine learning (ML) algorithms were implemented using the Keras library, while the classical methods were implemented using the Tensor Flow-Learn package. For this proposed research work, we gathered 54,306 pieces of data from weblogs, cookies, session usage, and from HTTP (S) request files to train and test our model. The performance evaluation results for training set in metrics such as the hybrid approach (ANN and SVM) perform better accuracies in precision (99.05% and 99.54%), recall (99.65% and 99.61%), f1-score (99.35% and 99.57%), and training set (99.20% and 99.60%) respectively than other ML approaches. However, their training time is too high (i.e., 19.62 and 26.16 s respectively) for NB and RF. Accordingly, the NB technique performs poorly in accuracy, precision, recall, f1-score, training set evaluation metrics, and best in training time. Additionally, the performance evaluation results for test set in metrics such as hybrid approach (ANN and SVM) perform better accuracies in precision (98.87% and 99.20%), recall (99.13% and 99.47%), f1-score (99.00% and 99.33%) and test set (98.70% and 99.40%) respectively than other ML approaches. However, their test time is too high (i.e., 11.76 and 15.33 ms respectively). Accordingly, the NB technique performs poorly in accuracy, precision, recall, f1-score, test set evaluation metrics, and best in training time. Here, among the implemented ML techniques, SVM and ANN are weak learners. The achieved performance evaluation results indicated that the proposed SQLI attack detection and prevention mechanism has been improved over the previously implemented techniques in the theme. Finally, in this paper, we aimed to keep researchers up-to-date, with contributions, and recommendations to the understanding of the intersection between SQLI attacks and prevention in the artificial intelligence (AI) field.

Calpito, Jomar & Olanday, Paul & Gallarde, Alain. (2022). **Application of advanced encryption standard in the computer or handheld online year-round registration system. Indonesian Journal of Electrical Engineering and Computer Science. 27. 922. 10.11591/ijeecs.v27.i2.pp922**-935. With various severe security threats for web applications, ensuring security on the database layer itself is imperative. Hence, this study aims to protect data saved on the computer or handheld online year-round (CHOY) registration system using the advanced encryption standard (AES) to strengthen data security within the app so that even potential attackers gain access to the app's database; they cannot obtain valuable information because it is scrambled and unreadable. The proponents based the study's conceptual framework on the symmetric and asymmetric key algorithms and procedures manual on enrollment of Southern Isabela College of arts and trades (SICAT) and ISO 25010. The study consists of three elements: developing the CHOY web app imbued with AES, testing it in terms of online registration and spam prevention, and evaluating it using the ISO 25010 in terms of compatibility, reliability, and security. The evaluation results show that implementing the AES in the CHOY web app meets the ISO 25010 criteria mentioned above.

Gizachew, Fitsum & Salau, Ayodeji & Hadi, Shaimaa & Mullu, Tsegay & Barud Demilie, Wubetu. (2022). **Development of a Compressive Framework Using Machine Learning Approaches for SQL Injection Attacks. Przeglad Elektrotechniczny. 7. 181-187. 10.15199/48.2022.07.30.** Web applications play an important role in our daily lives. Various Web applications are used to carry out billions of online transactions. Because of their widespread use, these applications are vulnerable to attacks. SQL injection is the most common attack, which accepts user input and runs queries in the backend and returns the desired results. Various approaches have been proposed to counter the SQL injection attack; however, the majority of them have most times failed to cover the entire scope of the problem. This research paper investigates the frequent SQL injection attack forms, their mechanisms, and a way of identifying them based on the SQL query's existence. In addition, we propose a comprehensive framework to determine the effectiveness of the proposed techniques in addressing a number of issues depending on the type of the attack, by using a hybrid (Statistic and dynamic) approach and machine learning. An extensive examination of the model based on a test set indicates that the Hybrid approach and ANN outperforms Naive Bayes, SVM, and Decision tree in terms of accuracy of classifying injected queries. However, with respect to web loading time during testing, Naive Bayes outperforms the other approaches. The proposed Method improved the accuracy of SQL injection attack prevention, according to the test findings.

Gizachew, Fitsum & Mullu, Tsegay & Demilie, Wubetu. (2022). **Attacks on SQL Injection and Developing Compressive Framework Using a Hybrid and Machine Learning Approach. 10.21203/rs.3.rs-1321852/v1.** Web applications play an important role in everyday life. Various Web applications are used to carry out billions of online transactions. These applications are vulnerable to attacks due to their extensive use. The most prevalent attack is SQL injection, which accepts user input and runs queries in the backend based on the user's input, returning desired results. To counter the SQL injection attack, various approaches have been offered; however, the majority of them either fail to cover the full breadth of the problem. This research paper looks into frequent SQL injection attack forms, their mechanisms, and a way of identifying them based on the SQL query's existence. We propose a comprehensive framework for determining the effectiveness of techniques that address certain issues following the essence of the attack, using hybrid (Statistic and dynamic) and machine learning. An extensive examination of the model based on a test set indicates that the Hybrid and ANN approaches outperform Naive Bayes, SVM, and Decision trees in terms of accuracy in classifying injected Queries. However, when it came to web loading time during testing, Nave Bayes outperformed. The Hybrid Method improved the accuracy of SQL injection attack prevention, according to the test findings. Although we used a limited dataset for training and testing in our study, it is advised that the dataset be expanded and the model be tested in a real-world setting.

Alanda, Alde & Satria, Deni & Ardhana, M.Isthofa & Dahlan, Andi & Mooduto, Hanriyawan. (2021). **Web Application Penetration Testing Using SQL Injection Attack. JOIV : International Journal on Informatics Visualization. 5. 320. 10.30630/joiv.5.3.470.** A web application is a very important requirement in the information and digitalization era. With the increasing use of the internet and the growing number of web applications, every web application requires an adequate security level to store information safely and avoid cyber attacks. Web applications go through rapid development phases with short turnaround times, challenging to eliminate vulnerabilities. The vulnerability on the web application can be analyzed using the penetration testing method. This research uses penetration testing with the black-box method to test web application security based on the list of most attacks on the Open Web Application Security Project (OWASP), namely SQL Injection. SQL injection allows attackers to obtain unrestricted access to the databases and potentially collecting sensitive information from databases. This research randomly tested several websites such as government, schools, and other commercial websites with several techniques of SQL injection attack. Testing was carried out on ten websites randomly by looking for gaps to test security using the SQL injection attack. The results of testing conducted 80% of the websites tested have a weakness against SQL injection attacks. Based on this research, SQL injection is still the most prevalent threat for web applications. Further research can explain detailed information about SQL injection with specific techniques and how to prevent this attack.

Al-Fawa'reh, Mohammad & Al-Fayoumi, Mustafa & Nashwan, Shadi & Fraihat, Salam. (2021). **Cyber threat intelligence using PCA-DNN model to detect abnormal network behavior. Egyptian Informatics Journal. 23. 10.1016/j.eij.2021.12.001.** Security issues are the most critical challenges facing new technologies associated with the internet of things (IoT), big data, and cloud computing. A secure and efficient intrusion detection system (IDS) is crucial to detect security threats. Existing IDSs are known to suffer from many problems, most notably the high rate of false positive alerts, the long time required to detect attacks, and the inability to detect zero-day attacks, which can ruin companies. The weakness of IDS backend engines costs companies time in the investigation process. This paper proposes and enhances IDS detection mechanisms via two processes: using a deep neural network (DNN) model with new features for threat detection based on two assumptions related to handling zero-day attacks, with low computing power and resources, and a comprehensive solution for detection by merging the DNN model and principle component analysis (PCA) to increase security and performance. The proposed detection mechanism combines DNN, PCA, statistical, and knowledge-based approaches to offer significantly greater efficiency than existing IDS, as indicated by analytical and software results. A simulation model is used with up-to-date web attacks, distributed denial of service (DDoS), denial of service (DoS), brute force, insider infiltration, Botnet, and Heartbleed attacks. The proposed detection techniques for large networks are analyzed and complexity in the design is avoided by reducing the number of DNN model layers, thus minimizing detection time delay and false positives, while increasing security against network attacks. Integrating the proposed DNN with PCA, an innovative contribution, introduces robust IDS to significantly improve the detection time delay and security performance. The proposed model showed a 98% accuracy rate. To best of our knowledge, the highest accuracy rate stated based on a large number of attacks is 97%, which makes our model state of art.

**CHAPTER THREE**

**SYSTEM ANALYSIS, DESIGN AND METHODOLOGY**

**3.1 OVERVIEW**

“Methodology is the philosophical framework within which the research is conducted or the foundation upon which the research is based” (Brown, 2006). The methodology of a research describes research methods, approaches and designs in detail highlighting those used throughout the study, justifying my choice through describing advantages and disadvantages of each approach and design taking into account their practical applicability to our research. The methodology should meet the following two criteria: Firstly, the methodology should be the most appropriate to achieve objectives of the research.  
Secondly, it should be made possible to replicate the methodology used in other researchers of the same nature (Randy, 2005).

The following methods were used to do extensive review on related works regarding the predictive system.

1. Internet: Browsing the internet, particularly Google Scholar and associated scholarly websites was used to get information regarding the predictive system.
2. Journals: Getting information from published journals on related works.
3. Gathering and analyzing of historical clinic data through the steps below.
4. Data requirement: Identifying and documenting the variables and entries within the scope of the project and preliminary data characteristics that will influence the system
5. Data Collection
6. Data Processing: This is the conversion of data into meaningful form.
7. Modelling: This is the development of a model which can be a programmed entry.

**3.2 ANALYSIS OF EXISTING SYSTEM**

### 3.2.1 Drawbacks of Manual Disease Prediction

Once a web application is exposed to SQL injection, a hacker can execute any malicious SQL inquiry or command through the web application. This means he or she can recover all the data saved in the database such as client information, social security numbers, credit card details and credential to access private areas of the gateway, such as the administrator portal. By utilizing an SQL injection, it is also likely to remove tables from the database. Hence, with an SQL Injection the malicious user has complete access to the database.

Depending on your setup and the kind of server software being used, by utilizing SQL injection vulnerability, some malicious users might also be able to write to a file or accomplish operating system commands. With such increased privileges this might result in a total server compromise.

Unfortunately, it is very hard to determine the impact of an exploited SQL injection. Most of the times, if the hackers are well trained, you won’t be able to identify the attack until your data is available to the public and your business reputation is going down the drain. (Oraf, 2019)

**3.3 PROPOSED MODEL**

Requirements for detecting web vulnerabilities in this project includes the use of ensemble learning that is combining five algorithms (Random Forest, Support Vector Machine, Decision Tree, K Nearest Neighbour, Naïve bayes algorithm). This is chosen because of its ability to analyse behaviours and modelling prediction to deal the web vulnerabilities and thereby making internet safer.

The Naïve Bayes algorithm performs well with categorical input variables but performs less well with numerical values and in multi-class classification/prediction. Additionally, the assumption of independence feature upon which the algorithm is based may not always be true.

The decision tree (DT) algorithm performs well for continuous as well as categorical variables. DT classifiers learn to make predictions on the test data by following a tree-like model (created using the training dataset) that resembles a flow chart, based on the features passed into it. Each of the tree’s internal nodes correlates with an attribute, and every leaf node correlates with a class label. In a DT, the best feature of the dataset is positioned at the root of the tree, while the training dataset is divided into subsets. These two steps are then repeated on every subset until there are no further divisions possible. It is a simple algorithm that can work well with large datasets.

Random forest (RF) is a classification algorithm that consists of multiple decision trees that make predictions based on the mean probabilistic prediction of each tree. It is similar to decision trees and it reduces the problem of overfitting, which is a problem associated with the DT algorithm. But it is not easy to interpret, unlike a decision tree. It uses randomness when constructing each DT to create a forest of different trees. Boosting is a method of making strong learners out of weak learners by combining weak classifiers into one strong classifier.

Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well its best suited for classification. The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. The dimension of the hyperplane depends upon the number of features. If the number of input features is two, then the hyperplane is just a line. If the number of input features is three, then the hyperplane becomes a 2-D plane. It becomes difficult to imagine when the number of features exceeds three.

K-nearest Neighbor is used to solve the classification model problems. K-nearest neighbor or K-NN algorithm basically creates an imaginary boundary to classify the data. When new data points come in, the algorithm will try to predict that to the nearest of the boundary line.

Therefore, larger k value means smother curves of separation resulting in less complex models. Whereas, smaller k value tends to overfit the data and resulting in complex models.

**3.3.1 Benefits Of The Proposed Model**

1. It helps reduces cases of fraud: Most times there is a breach sensitive information’s are exposed to the hands of fraudulent individuals and it could be used for various incessant activities this project helps defeat that act.
2. Increase in Sales: Companies that have their websites crashed due to sql injection tend to lose customers but those that can detect this can easily make plans to stop it from happening thereby increasing sales in the process.

**3.4 Data Set**

A data set is a collection of related, discrete items of related data that may be accessed individual or in combination or managed as a whole entity. Machine learning usually works with two data sets: training and testing data.

**3.4.1 Training Data**

Training data is an initial set of data used to help a program understand how to apply technology like neural networks to learn and produce sophisticated results.

Trained Data is labeled data used to train machine learning algorithms and increase accuracy. The observation in the training set from the experience that the algorithm uses to learn.

**3.4.2 Test Data**

The test set is a set of observations used to evaluate the performance of the model using some performance metric. It is important that no observation from the training set are included in the test set. If the test set does contain examples from the training set, it will be difficult to assess whether the algorithm has learned to generalize from the training set or has simply memorized it.

A program that generalizes well will be able to effectively perform a task with new data. In contrast, a program that memorizes the training data by learning an overly complex model could predict the value of the response variable for the training set accurately, but will fail to predict the value of the response variable for new examples. Memorizing the training set is called over fitting. A program that memorizes its observations may not perform its task well, as it could memorize relations and structures that are noise or coincidence. Balancing memorization and generalization, or over-fitting and under-fitting, is a problem common to many machine learning algorithms. Regularizations may be applied to many models to reduce over-fitting.

**3.4.3 Data Pre-processing**

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues.When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data preprocessing task.

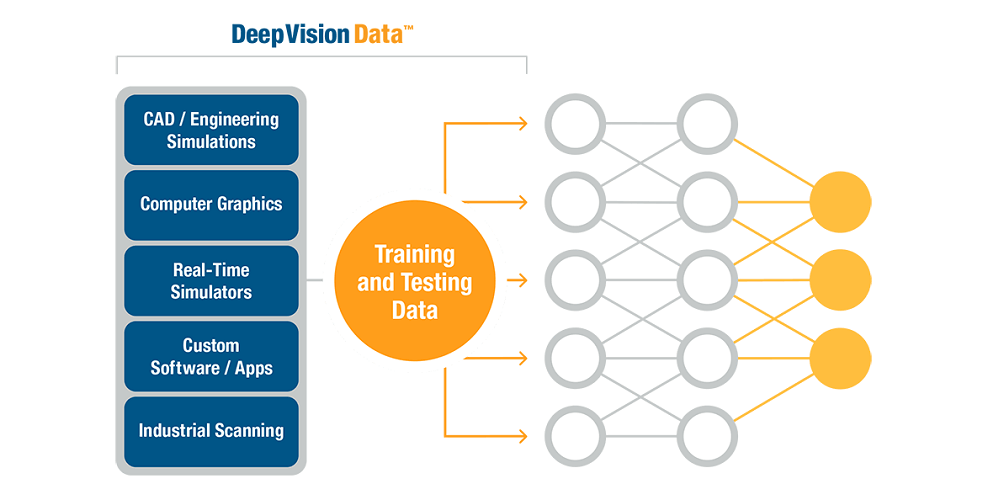
**3.4.4 Feature Selection**

Feature selection is a process of dimensionality reduction by which an initial set or raw data is reduced to more manageable groups for processing. A characteristic of these large data sets is a large number of variables that require a lot of computing resources to process. Feature selection is the name for methods that select and/or combine variables into features, effectively reducing the amount of data that must be processed, while still accurately and completely describing the orginal data set (deepai 2017).

The process of features selection is useful when you need to reduce the number of resources needed for processing without losing important or relevant information. Feature selection can also reduce the amount of redundant data for a given analysis, Also, the reduction of the data and the machine's effort in building variable combinations (features) facilities the speed of learning and generalization steps in the machine learning process.

**3.4.5 Training algorithm**

Training algorithm will involve training an algorithm or machine learning model to predict the outcome you design your model to predict.



**Figure 3.1 Training data**

**3.4.6 Predictive model**

Predictive modelling is the general concept of building a model that is capable of making predictions. Typically, such a model includes a machine learning algorithm that learns certain properties from a training dataset in order to make those predictions.

**3.4.7 Supervised learning**

Supervised learning is the machine learning task of leaarning a function that maps an input to an output based on example input-output pairs (Russel & Norvig, 2010). It infers a function from labeled training data consisting of a set of training examples (Mohri et al, 2012). In supervised learning each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances. This requires the learning algorithm to generalise from the training data to unseen situation in a reasonable way.

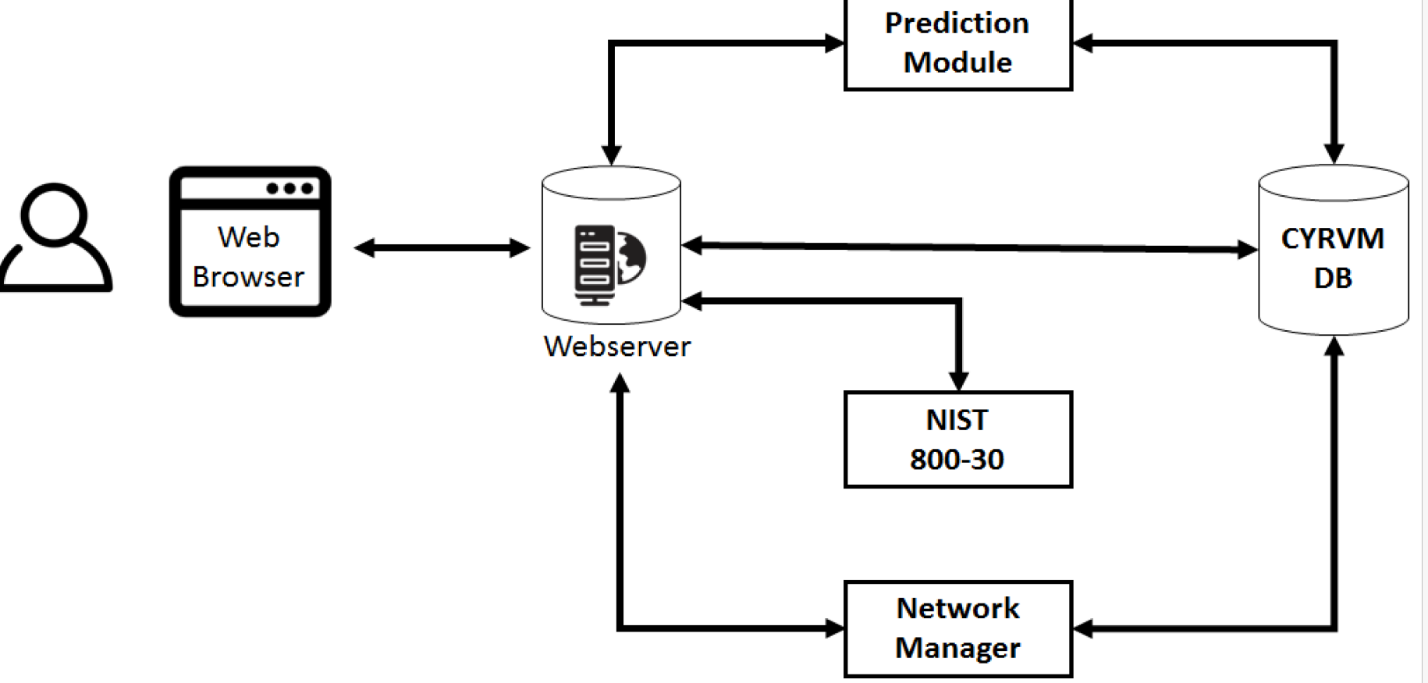
**3.4.8 Unsupervised learning**

Unsupervised learning is a machine learning technique, where you do not need to supervise the model. Instead, you need to allow the model to work on its own to discover information. It mainly deals with the unlabeled data. Unsupervised learning algorithm allows you to perform more complex processes tasks compared to supervised learning. A central application of unsupervised learning is in the field of denisty estimation in statistics (Jordan, 2004).

## **3.5 System Implementation**

### 3.5.1 Architecture of the System

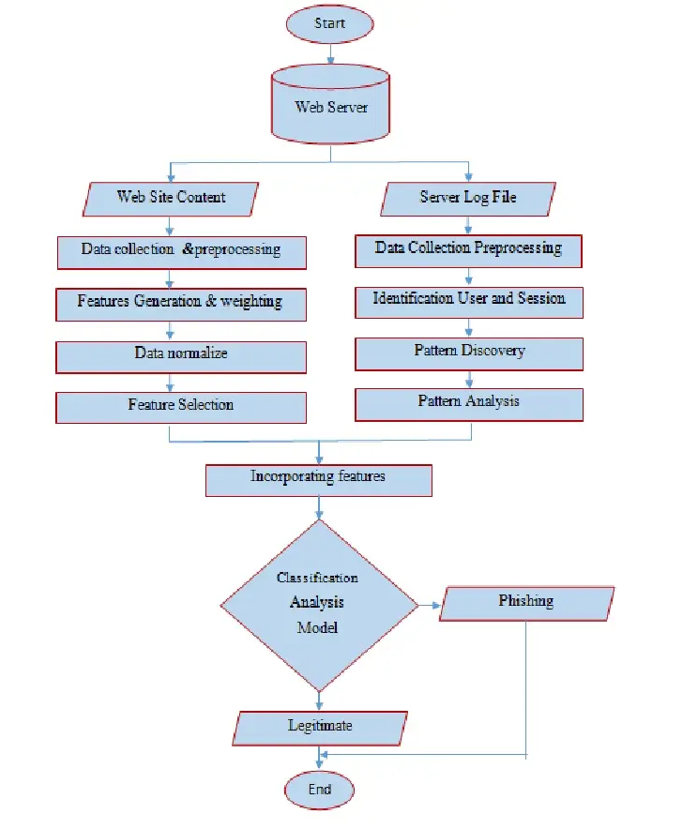
The system architecture is a 2- tier system where the first level runs the various machine learning models for web vulnerability, predicts their detection and the second level uses an insertion sort algorithm to return the attack with the highest probability to the user/worker. As shown in the architectural diagram below, the system takes the user’s biodata as input and uses it to run all the machine learning models. The results are then saved in the database then sorted, and the most likely attack is displayed for the user.



**Figure 3.2 Architecture of the system**

### 3.6 Flowchart Of The System

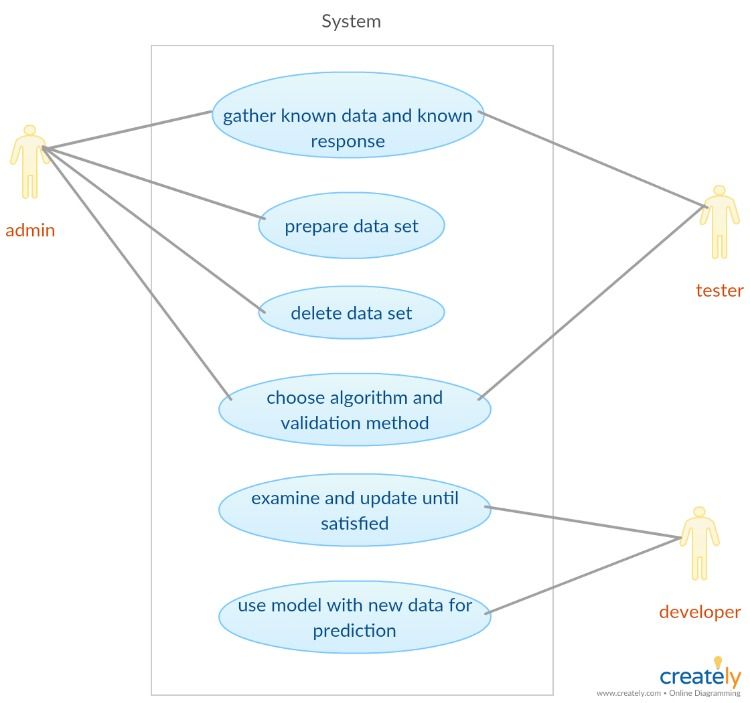
The flowchart of the system shows the steps involved in its execution. It thoroughly shows the process flow from accepting biodata as input and running the model to displaying the result to the user. Each step in the process is represented by a different symbol and contains a short description of the process step. The flowchart symbols are linked together with arrows showing process flow direction.



**Figure 3.3 Flow chart of the system**

### 3.6.1 Use Case Diagram Of The System

The use case diagram is a diagrammatic representation of how the user interacts with the system proposed. It shows all interactions or actions that are possible by the actors (Workers and Administrators) of associated with the system.



## **3.7 System Testing**

The developed system will be tested thoroughly by using test data.