Pavithra S. International Journal of Engineering Research and Applications
www.ijera.com
ISSN: 2248-9622, Vol. 13, Issue 10, October 2023, pp 61-67

RESEARCH ARTICLE OPEN ACCESS

An Approach Using Machine Learning to Predict Thyroid Disease in the Early Diagnosis Stages: A Review

Pavithra S
Lecturer
Department of Electronics and Communication Engineering/Government Polytechnic Udupi
pavithra.innanje@gmail.com

ABSTRACT
The thyroid gland's edge experiences an abnormal proliferation of thyroid tissue, which causes thyroid illness. The two primary forms of thyroid disorders are hypothyroidism (an inactive thyroid gland) and hyperthyroidism (an overactive thyroid gland), which often result when this gland generates excessive amounts of hormones. To identify and diagnose thyroid illness, this study recommends the usage of effective classifiers by utilizing machine learning algorithms in terms of accuracy and other performance assessment criteria. In this study, a wide range of classifiers, including K-nearest neighbor (KNN), Naive Bayes, support vector machines, decision trees, and logistic regression implemented with or without feature selection methods, are extensively analyzed. The thyroid dataset included the three extra parameters of pulse rate, body mass index, and blood pressure, it stood out from previous research already conducted. The experiment consisted of three iterations; the first iteration did not use feature selection, while the second and third used a feature selection approach based on L1 and L2.
Numerous aspects of the experiment have been evaluated and analyzed, including accuracy, precision, and the receiver operating curve's area under the curve. The outcome showed that classifiers using L1-based feature selection outperformed classifiers using L2-based feature selection in terms of overall accuracy (Naive Bayes 100%, logistic regression 100%, and KNN 97.84%).
Keywords: KNN, CNN, Thyroid Disease, SVM, Feature Selection, Deep Learning, EDA, NB, Machine Learning Techniques, Classification.

I. INTRODUCTION

It is difficult to diagnose thyroid illness. It calls for several steps. The typical, conventional approach is a thorough physical examination and several blood samples for blood testing. Therefore, a model that recognizes thyroid illness at a very early stage of development is required [1]. Machine learning has various classification models on which we can train our model with appropriate training datasets of thyroid patients and can predict and give the results in an accurate manner with higher degree of correctness, which is important in the medical field for thyroid disease diagnosis. Machine learning has various classification models on which we can train our model with appropriate training datasets of thyroid patients and can predict and give the results in an accurate manner with higher degree of correctness, which is important in the medical field for thyroid disease diagnosis.

The medical team may quickly assess the patient's condition based on the test results, and if required, forego additional clinical investigations. Thus, the healthcare industry benefits greatly from this strategy. A suitable train dataset produces a reliable forecasting model, which lowers the total cost and shortens the treatment time for thyroid patients [2]. The best algorithms for making decisions and resolving issues in the real world are classification algorithms.

In addition to the clinical and crucial examination, accurate interpretation of thyroid illness is crucial for the purposes of diagnosis. The authors Chen et al. [11] discuss the significance of feature selection approach for enhancing classification accuracy helpful for purposes of diagnosis. With the use of the L1 and L2 feature selection techniques, the efficacy of several classification methods was examined in this research. As a result, it is assumed that newly incorporated features would offer precise and reliable measurements for detecting thyroid condition.

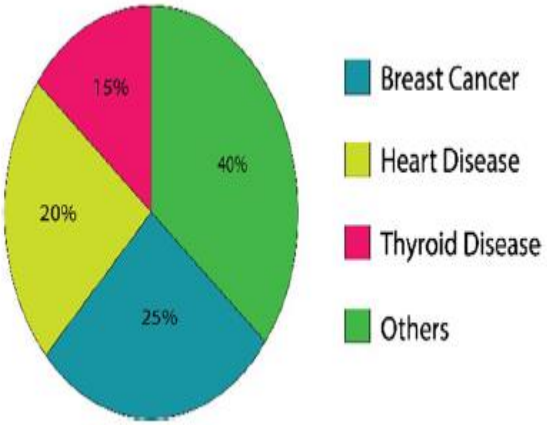

Figure No. 1: Health care statistics using classification.

Unique thyroid dataset was used in the investigation. The effectiveness of the proposed research is assessed using the confusion matrix, and the acquired findings were also compared with previous studies that concentrated on thyroid diagnosis and are listed in Table 3. The overall structure of the paper is as follows: A literature review and a dataset are included in sections 1 and 2, respectively. Section 3 describes in full the approach used. Analyses of the experimental data are presented in Sections 4 and 5. In Section 6, we highlight relevant recent studies. The paper's future scope is discussed in Section 7.

II. RELATED WORK

The methodology used in this investigation has been compared to other relevant, recent studies, which are listed in Table. Due to three novel properties discussed in Section 2, our model dataset differs from this previous research. The suggested study's findings were attained by employing several supervised classifiers. The important objectives of this research were to increase accuracy while minimizing training and forecast times. Other existing models employ hybrid strategies that combine various algorithms and intricate models. Such approaches take more time for training and validation and are more expensive to use to obtain reliable data.

Using data mining techniques, ***Bibi Amina Begum et al. [1]*** have presented various Thyroid prediction strategies. They discussed classification approaches in data mining such Decision Tree, Backpropagation Neural Network, SVM, and density-based clustering. They also considered various dataset features for prediction. They have examined the relationship between T3, T4, and TSH and hyper- and hypothyroidism.

Machine learning techniques based on classification have been researched by ***Ankita Tyagi et al. [2].*** They evaluated and assessed the performance metrics of the decision tree, support vector machine, and K-nearest neighbor algorithms using the train data set from the UCI Machine Learning repository. A training model with 21 thyroid-causing features has been proposed by ***Aswathi A. K. et al. [3].*** To improve the settings of the support vector machine, they suggested using partial swarm optimization.
In a broad empirical investigation on the diagnosis of several diseases, including diabetes, breast cancer, heart disease, and thyroid prediction, ***M. Deepika et al. [4]*** examined the accuracy rate using SVM, decision trees, and artificial neural networks. Thyroid data preparation was primarily studied by ***Sumathi A et al. [5]*** by using the decision tree approach. The preprocessing stage was started by calculating the mean values of T3, T4, and TSH. Later, they used machine learning-based feature building and selection techniques. Additionally, they used the classification based J48 method, a development of the ID3 algorithm, and computed the outcomes.

III. DATA DESCRIBTION

Hypothyroidism is the most prevalent thyroid condition. Hypothyroidism is a disorder in which the thyroid gland is underactive or produces too little thyroid hormone. Hypo- denotes deficient or under(active). It is crucial to recognize the signs of hypothyroidism.

Figure No. 1 – Thyroid Conditions

In our paper, we have considered 6 databases from the Garavan Institute in Sydney, Australia
Approximately the following for each database:

- 2800 training (data) instances and 972 test instances.
- Plenty of missing data
- 29 or so attributes, either Boolean or continuously valued.

Two additional databases, also from Ross Quinlan, are also considered.

- Hypothyroid.data and sick-euthyroid.data

- Quinlan believes that these databases have been corrupted.
- Their format is highly like the other databases.

A Thyroid database suited for training ANNs and CNNs Algorithm

- 3 classes
- 3772 training instances, 3428 testing instances.
- Includes cost data (donated by Peter Turney)

IV. PROPOSED METHODOLOGY

The methodology described in this work includes a few crucial phases, which are shown in Fig. 1. The first phase in our technique is data processing, which comprises clearing out any unnecessary columns or items. Cleaning out superfluous data and processing missing values might potentially increase overall result correctness. Furthermore, handling missing values is essential since omitting them might result in the loss of important data, which would have a detrimental effect on the outcomes. This is followed by the implementation of feature scaling using the min-max approach to determine the highest and lowest entry values. The initial portion of the experiment is carried out without the use of feature selection approaches in order to get efficient accuracy and performance of the classifiers. The experiment is divided into three phases, with the second and third phases implementing L1- and L2-based feature selection algorithms, respectively. This study includes features like blood pressure, pulse rate, and BMI since they have a direct correlation with thyroid diseases and are essential for getting the most accurate findings. Different evaluation metrics, including the f1-score, miss-rate, Matthew correlation coefficient (MCC), error-rate, ROC curve with area under the curve (AUC), sensitivity, selectivity, fall-out, and accuracy, have been used to assess and compare the various classifiers and top algorithms for identifying thyroid disease.

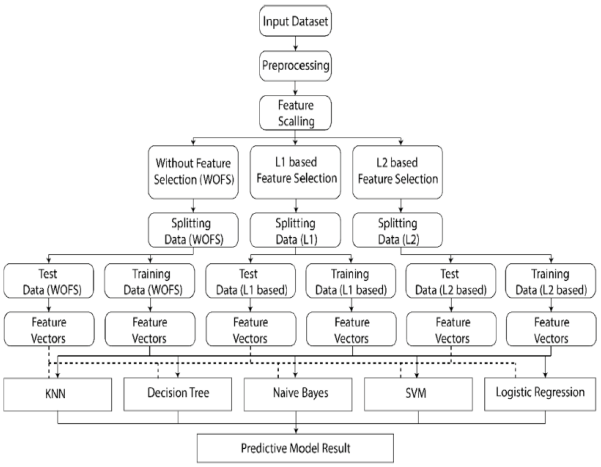

Figure No. 2 : Proposed Methodology

4.1: FEATURE SELECTION

To improve a given classifier's effectiveness, feature selection is essential. The millions of pieces of data that modern IoT devices deliver result in datasets with hundreds of undesirable attributes. As a result, these traits stifle the model, exponentially lengthening training, and raise the danger of overfitting. With the help of the feature selection approach, training and prediction may be done in less time on average without losing any information. In order to save money and time, these key characteristics were later employed for testing and training. These methods significantly affect the categorization outcomes [12].

4.2: NORMAL BASED FEATURE SELECTION

With the aid of the scikit-learn Python package, the L1 and L2 based feature selection method has been applied for this study. Scikit-learn is a very user-friendly library with a surprising reaction speed of different algorithms and approaches, compared to other existing libraries as mlpy, pybrain, and shogun [25]. To accomplish dimensionality reduction for specific datasets, classifiers can be employed in conjunction with these L1- and L2-based feature selection algorithms. Some coefficients are given a value of zero by L1 feature selection approaches. Because they don't affect the final forecast, certain traits are eliminated as a result of target estimate. However, the L2 feature selection strategy approaches zero rather than assigning the coefficient value zero. The linear support vector classifier (LSVC) was employed for this study, and a C parameter was chosen to regulate the sparsity. Observation reveals that the value of C and the number of characteristics chosen are closely correlated; the higher the value of C, the more features chosen; and vice versa.

V. DIFFERENT APPROACH

5.1: KNN

The most popular and commonly used supervised machine learning method is K-nearest neighbors (KNN). For predictive analysis and pattern identification, KNN works well. Predicting discrete values in classification problems is one of the key applications of KNN [7]. KNN functions as a classifier using the distance function or similarity measure and the chosen k value, with the performance relying on the criteria. KNN first determines the distance between each data point and each new data point, then it accumulates the ones that are nearby.

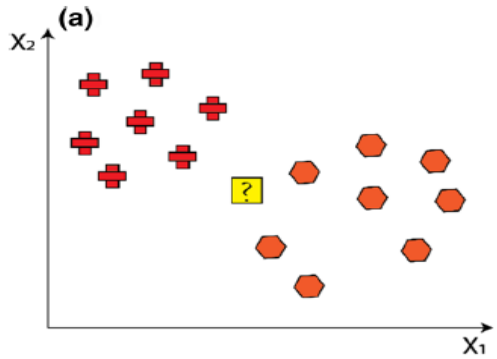

Figure No. 3: Working of KNN Method

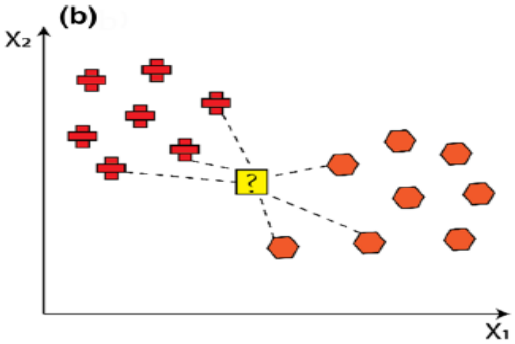

Figure No. 4: Initial Data

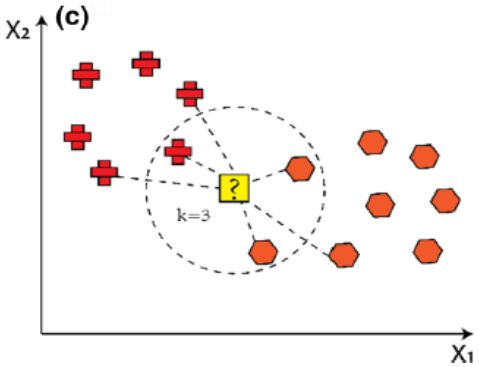

Figure No. 5 : Finding Neighbor and Vote

The nearest data points are then arranged by the algorithm according to how far they are from the arrival data point using various distance functions. The following phase is gathering a particular number of the data points with the shortest distance among all of them and classifying them according to that distance. The operation of KNN is shown in Figure 3. The red plus symbol in the illustration belongs to class 01, whereas the green plus sign is from class 2. The method would predict that the yellow box point "?" on the diagram is either associated with class01 or class02, respectively. Assuming feature vectors a and a = a1, a2,…an and b = b1, b2,…bn The following is a discussion of the distance functions under consideration:

$$\cos ine_{d(a,b)} = \frac{\sum_{i=1}^{n} (a_i)(b_i)}{\sqrt{\sum_{i=1}^{n}} (a_i)^2 \sqrt{\sum_{i=1}^{n}} (b_i)^2}$$

$$canberra_{d(a,b)} = \sum_{i=1}^{n} \left(\frac{|a_i - b_i|}{|a_i| + |b_i|}\right)$$

5.2. SUPPORT VECTOR MACHINE (SVM)

A supervised machine learning approach called the support vector machine (SVM) may be used to do classification, regression, and even outlier identification. The dataset's characteristics are represented graphically in n-dimensional space. By sketching a straight line known as a hyperplane, the two classes may be distinguished [9]. The points in the dataset that fall on one side of the line will all be categorized as belonging to the same class, while the points that fall on the other side of the line will all be given a second-class designation. The tactic seems straightforward enough, but it's vital to remember that there are an endless number of lines available. SVM aids in choosing the line that categorizes the data the most accurately. In addition to choosing a line to draw between the two classes, the SVM algorithm also avoids getting too close to the nearby samples. The "support vector" in the phrase "support vector machine" really refers to two position vectors drawn from the origin to the points that define the decision boundary [3]. Figure 4 depicts the SVM's operating system.

5.3. NAÏVE BAYES CLASSIFICATION

Naive Bayes is a straightforward technique for classifying different classification issues. It is simple to construct and has a great deal of predictive power and accuracy. This classifier uses the Bayesian theorem-based probabilistic learning approach [1]. The three steps are the basis for the operating principle. The dataset is initially transformed into a frequency table. The likelihood table is created in the second stage after the probabilities are determined. The Naive Bayes

equation is used in the last stage to determine the posterior probability for each class. The result of a prediction is the class with the highest posterior probability rate [30]. Bayes's theorem is thus as follows:

$$P\left(\frac{h}{D}\right) = \frac{\left[\left(P\frac{D}{h}\right) \cdot P(h)\right]}{P(D)}.$$

5.4. DECISION TREE

Decision trees are a well-known technique for making decisions. By partitioning the instance space into decision zones, a novel "divide-and-conquer" tactic is applied. A root node is developed after some testing. The value of the relevant test property is then used to split the dataset.

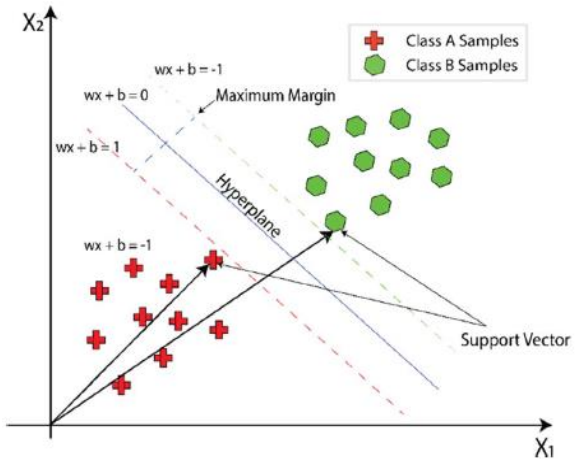

Figure No. 6: Decision Tree Technique

By offering a predetermined stopping condition, the process of repetition may be stopped. A leaf node, which is a node at the end of a tree, serves as a class indicator. The branch or route of the node defines the decision rule. Every fresh sample has a different categorization decision rule [10]. There are three phases involved in these classifications. First, the model is trained during the learning process using training data. Second, a test is run to determine the model's correctness, and depending on the result, the model is either approved or rejected. An accurate and widely accepted value is required to apply the model for further categorization of a new datum. Thirdly, and most importantly, the model's application is determined by applying it to forecast new data or classify existing data [9]. While the workings of a decision tree are seen in Fig. 5, the definitions of entropy and the gini equation are given below in equation.

$$\text{Ent}(D) = -\sum_{y \in Y} P(y|D) \log P(y|D).$$

$$G_{\text{gini}}(D; D_1 \ldots, D_k) = I(D) - \sum_{i=1}^{k} \frac{|D_k|}{D} I(D_k)$$

$$\text{where } I(D) = 1 - \sum_{y \in Y} P(y|D)^2.$$

VI. ACCURACY AND ERROR

Accuracy is the most significant and often used criterion to assess the performance of the classifier. The ratio of correct prediction samples to all samples in the dataset is used to compute accuracy (ACC).

$$\text{Accuracy(Acc)} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100\%$$

However, error rate (ERR), which is determined as follows, indicates the number of samples that were incorrectly categorized into the negative and positive classes.

$$\text{Error Rate} = (1 - \text{Acc}) \times 100\% = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100\%$$

VII. EXPERIMENTAL DATA

This research study's experimental examination included dependent on the system's hardware and software performance. The hardware setup utilized in this experiment included an Intel(R) Core i7-7700HQ CPU running at 2.80 GHz, a 512 GB SSD, a 2 TB HDD, 16 GB of RAM, and an Nvidia GTX 1060 GPU with 6 GB of RAM. On the other hand, the package description mentioned using Anaconda [38] and Scikit-Learn [25]. Because of its accessibility, simplicity, and excellent performance for data analysis, Scikit-learn was a fantastic pick. Five machine learning algorithms—KNN, decision tree, SVM, logistic regression, and Naive Bayes—were utilized, and they were split into training and testing datasets. The five machine learning algorithms employed the well-known L1 and L2 feature selection methods. Three times the thyroid dataset experiment was conducted. Without feature selection, or (WOFS), was the initial iteration. In the experiment's second try, L1-based feature selection, also known as WLSVC(L1), was used. With L2-based feature selection applied, the experiment's third iteration was known as WLSVC(L2).

If the classifier makes use of crucial characteristics, training and testing time can be cut down. The significance of feature selection is

dependent on several factors, one of which is the F-score, which assesses the value and significance of different characteristics. Regression and classification issues can be solved more effectively with the Xgboost classifier. This method uses less time and computational resources while prioritizing greater results. To stop the model from overfitting is the major goal of using Xgboost in L1 and L2 feature selection techniques. This study also used the xgboost classifier [39], which automatically identified the features according to their index in the input array, to choose different features based on their F-score values. Figure 7 shows the outcome of an experiment utilizing WOFS in which five characteristics were given weight by the algorithms. The findings show that f0 (TSH) is most significant and f4 (pulse rate) is least significant. Like this, four crucial characteristics were chosen for WLSVC(L1) implementation in Fig. 8b based on their F-scores. Where f0 (TSH) has the most relevance and f3 (BMI) has the lowest index in WLSVC(L1).

VIII. CONCLUSION

Early disease diagnosis and identification are crucial for maintaining human life. Precise and accurate identification and detection have become easier to achieve with the use of machine learning algorithms. Due to the symptoms of thyroid illness being confused with those of other conditions, diagnosis is difficult. The three newly added characteristics in the thyroid dataset in this study have a beneficial influence on classifier performance, and the findings demonstrate that they outperform previous studies in terms of accuracy. After comparing and analyzing KNN, Nave Bayes, SVM, decision trees, and logistic regression, it was found that Nave Bayes achieved 100% accuracy in each of the three experiment parts, while logistic regression achieved 100% and 98.92% accuracy in L1- and L2-based feature selection, respectively. KNN also produced great results, with a 97.84% accuracy rate and a 2.16% error rate. The benefits and resilience of the new dataset are evident after analysis and would enable clinicians to obtain more exact and accurate findings in less time. For more accurate findings, classifiers using other KNN distance functions and data augmentation approaches may be utilized in the future.

REFERENCE

[1]. Bibi Amina Begum and Dr.Parkavi "Prediction of thyroid Disease Using Data Mining Techniques" 5Th International Conference on Advanced Computing & Communication Systems (ICACCS), 2019

[2]. Ankith Tyagi, Ritika Mehra, Aditya Saxena "Interactive Thyroid Disease Prediction System Using Machine Learning Technique" 5th IEEE International Conference on Parallel, Distributed and Grid Computing (PDGC-2018), 20-22 Dec, 2018, Solan, India

[3]. Aswathi A K and Anil Antony "An Intelligent System for Thyroid Disease Classification and Diagnosis" Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018) IEEE Xplore Compliant - Part Number: CFP18BAC-ART; ISBN:978-1-5386-1974-2

[4]. M Deepika and Dr. K. Kalaiselvi "An Empirical study on Disease Diagnosis using Data Mining Techniques." Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018) IEEE Xplore Compliant - Part Number: CFP18BAC-ART; ISBN:978-1-5386-1974-2

[5]. Sumathi A, Nithya G and Meganathan S "Classification of Thyroid Disease using Data Mining Techniques" International Journal of Pure and Applied Mathematics, Volume 119 No. 12 2018, 13881-13890

[6]. Md. Dendi Maysanjaya, Hanung Adi Nugroho and Noor Akhmad Setiawan "A Comparison of Classification Methods on Diagnosis of Thyroid Diseases" 2015 International Seminar on Intelligent Technology and Its Applications

[7]. Shroff, S.; Pise, S.; Chalekar, P.; Panicker, S.S.: Thyroid disease diagnosis: a survey. In: IEEE 9th International Conference on Intelligent Systems and Control, 2015 (ISCO 2015), pp. 1–6. IEEE (2022)

[8]. Chandel, K.; Kunwar, V.; Sabitha, S.; Choudhury, T.; Mukherjee, S.: A comparative study on thyroid disease detection using K-nearest neighbor and Naive Bayes classification techniques. CSI Trans. 4(2–4), 313–319 (2020)

[9]. A. Colubri, T. Silver, T. Fradet, K. Retzepi, B. Fry, P. Sabeti, Transformingclinicaldata into actionable prognosis models: machine learning Framework and fielddeployableapp to predict outcome of Ebola Patients, PLoSNegl. Trop. Dis. 10 (3) (2016) e0004549.

[10]. Md. Dendi Maysanjaya, Hanung Adi Nugroho and Noor Akhmad Setiawan "A Comparison of Classification Methods on Diagnosis of Thyroid Diseases" 2015 International Seminar on Intelligent Technology and Its Applications

[11]. Bekar, E.T.; Ulutagay, G.; Kantarcı, S.: Classification of thyroid disease by using data

mining models: a comparison of decision tree algorithms. Oxf. J. Intell. Decis. Data Sci. 2016(2), 13–28 (2021)

[12]. Roshan Banu D and K.C.Sharmili "A Study of Data Mining Techniques to Detect Thyroid Disease" International Journal of Innovative Research in Science, Engineering and Technology (Vol. 6, Special Issue 11, September 2020)

[13]. Dr. Srinivasan B, K.Pavya "Diagnosis of Thyroid Disease Using Data Mining Techniques: A Study" International Research Journal of Engineering and Technology Volume: 03 Issue: 11 | Nov - 2019

[14]. SunilaGodara and Sanjeev Kumar "Prediction of Thyroid Disease Using Machine learning Techniques" International Journal of Electronics Engineering (ISSN: 0973-7383) Volume 10 • Issue 2 pp. 787-793 June 2018

[15]. Ali keles et al.,"ESTDD: Expert system for thyroid diseases diagnosis", Expert system with Applications,34,242-246,200