How to Cite:
Veeranki, S. R., & Varshney, M. (2022). Comparative analysis of thyroid disease and predict them using machine learning techniques. *International Journal of Health Sciences*, *6*(S3), 11005–11014. https://doi.org/10.53730/ijhs.v6nS3.8459

Comparative analysis of thyroid disease and predict them using machine learning techniques

Sreenivasa Rao Veeranki
Department of Computer Science and Engineering, School of Engg. & Tech., Maharishi university of Information Technology, Lucknow, India
Email: sreeni.bi@gmail.com

Manish Varshney
Department of Computer Science and Engineering, School of Engg. & Tech., Maharishi university of Information Technology, Lucknow, India
Email: itsmanishvarshney@gmail.com

***Abstract*---**Bioinformatics is the field of research where the computational process has been used to analyse biological information. Genetic structure of a living creature decides the characteristics of the creature. By analysing the genetic structure, the details about the living creature can be known. Any disease occurred, is generated by the effect of any microorganisms. The genetic structure of the microorganisms has to know to deal with the microorganisms. The genetic structure of the microorganisms is very much useful to discover drug to protect the disease caused by the microorganisms. Similarly, in human body, some disease occurred due to the abnormal functionality of the human organs. These abnormal functionalities have been caused by the disputes in the genetic structure of the human beings. Thyroid is a disease caused by the abnormal functionality of the Thyroid gland. Bioinformatics based research work helps to identify the genetic structures of human being that is responsible for the thyroid disease. In this work, Bioinformatics technique has been applied to deal with the Thyroid disease. A public dataset generated by the Garavan Institute has been used for the research work. Three machine learning based classifiers Random Forest, k Nearest Neighbour, and Support Vector Machine have been used in this work to classify the thyroid disease data.

***Keywords*---**thyroid disease, bioinformatics, machine learning, random forest, k-nearest neighbour, support vector machine.

11006

Introduction

Thyroid is an endocrine gland which is located in the neck in a position outside of the trachea and just beneath the Adam's apple. The thyroid gland consists of two lobes, such as, one is right lobe which is presented at slightly right side of the trachea and another one is left lobe which is situated at slightly left side of the trachea. These two lobes of thyroid gland are connected by a small bridge which is called the isthmus. As the thyroid gland is a part of the endocrine system, it acts as messenger by secreting chemical factors which are called hormones to make a feedback loop to connect as well as regulating the distant organs of the body. The endocrine glands are those kinds of glands which secret hormones. These secreted hormones directly released into blood stream and travel with the blood to the tissues of the target organs which are situated at near or distant portion of the body in respect to secreted sight or glands. There are many types of endocrine glands present in our body to regulate our organ functions properly. Some endocrine glands are thyroid gland, adrenal gland pituitary gland, pineal gland, parathyroid gland, pancreas, ovaries (in female) and testes (in male) etc.

The thyroid gland is one of the most important endocrine glands which help in regulating growth and development of the health. Thyroid gland secrets two thyroid hormones, such as, triiodothyronine (T3), tetraiodothyronine or thyroxine (T4) and calcitonin. The biochemical formula of T3 is L-3,5,3'-triiodothyronine and T4 is L-3,5,3',5'-tetraiodothyronine. The predominant form of thyroid hormone in the blood is thyroxine or T4, T4 has a longer half-life than T3 (Irizarry, 2014). Thyroid glands store these thyroid hormones and release them at the time of requirement to play their function in needed organs. Thyroid gland is controlled by the two vital glands present in brain, such as hypothalamus and pituitary gland. The hypothalamus secretes thyrotropin releasing hormone (TRH) which helps in stimulating the pituitary gland to release thyroid stimulating hormone (TSH). The main component of these two thyroid hormones, T3 and T4, is iodine. Thyroid gland uptake iodine from foods which we intake, to produce these two main hormones, T3 and T4. So, iodine (I) is an essential inorganic nutrient which should be included in our diet. Presence of insufficient amount of iodine in our body leads to hamper in the making process of thyroid hormones. If we intake iodinated foods in excess level then it also gives effect on the health which should be avoided. Excessive as well as insufficient iodine intake may create diseases. Function of thyroid hormones: Effect of thyroid hormones on the cells of the tissues of all organs is as follows:

- The thyroid hormones help to burn the calories through hormonal actions resulting that, weight loss or weight gain may occur.
- They help in regulation of carbohydrate, protein and fat metabolism.
- They help in proper development as well as participate in differentiation of all cells of the body.
- The thyroid hormones help in neuronal and glial cell differentiation, growth and development and migration, synaptogenesis and myelination.
- Thyroid hormones help to proliferate the cells of bones and differentiating cellular functions of the bones such as chondrocytes, osteoblasts and osteoclasts (Kim, 2013).

- Thyroid hormones help in regulating many growth factors such as fibroblast growth factor, parathyroid hormones, insulin like growth factor-1, Indian hedgehog and Wnt to control skeletal muscles.
- Calcitonin plays a crucial role in regulating calcium and phosphate levels in the blood which is important for the bone health and maintenance.

Thyroid Hormones Abnormalities
Hyperthyroidism

The phenomena of producing excessive amount of thyroid hormones are referred to as hyperthyroidism. These following conditions are the reasons of causing hyperthyroidism:

- ***Graves' Disease:*** In this disease, thyroid gland is enlarged. The enlargement of the thyroid glands is called goitre. In this case, thyroid hormones become over activated and as a result of that, excessive amount of thyroid hormones are produced.
- ***Excessive Iodine Intake:*** The too much amount of iodine intake through foods can leads to produce excessive amount of the thyroid hormones than the requirement.

Hypothyroidism

The phenomena of producing little amount of thyroid hormones is called hypothyroidism. These following conditions are the reasons of causing hypothyroidism:

- ***Thyroiditis:*** These disease leads to the inflammation of the thyroid gland. This disease can decrease the amount of thyroid hormones production.
- ***Hashimoto's thyroiditis:*** This is an auto-immune disease. Auto-immune disease generally occurred due to genetic abnormalities. In this case, cells of the immune system of a body can inhibit and destroy the thyroid glands of itself due to the immune cells treat them as a foreign particle.
- ***Iodine Deficiency:*** Iodine is a crucial factor for the thyroid hormones (T3 and T4) production. Due to Iodine deficiency, body cannot make proper amount of thyroid hormones.

Analysis by bioinformatics to cure the disease related to thyroid hormone deficiency:

- Bioinformatics analysis is utilized to find the crucial genes as well as various types of pathways.
- By the help of bioinformatics, many genes can be identified.
- There have some specific tools or databases for identifying the genes, such as,
  - GenBank database is used for identifying new sequences of de-oxy-ribo-nucleic-acids (DNA).
  - For annotation purpose, Visualization and Integrated Discovery (DAVID).
  - For the identification of the interaction between proteins, searching tool for retrieval interacting genes (STRING) 10.0.

- By the help of bioinformatics through using tools and databases, scientists can find cell surface receptor linked signal transduction, leukocyte response, cytokine-cytokine interaction, interferon-gamma actions, different immune responses at various biological state, inflammatory responses, natural killer cell-mediated cytotoxicity, antigen processing and antigen presentation procedure etc (Zheng, 2020).
- Through the help of bioinformatics, it can be explored that, if there are any possibilities of graft rejection in hosts.
- Different algorithms as well as tools and databases of the bio-informatics can predict as well as make suitable designs of drugs which helps in the drugs discoveries also.

Bioinformatics have immense role in biological sciences which should be more explored by the researchers.

Literature Review

Bioinformatics tools have been applied to handle thyroid patients in recent past quite successfully. In the work (Zhi-long, 2012), the authors have tried to identify biomarkers, which are very much reliable to predict the disabilities of the thyroid gland. The identified biomarkers are helpful to predict the thyroid nodules. The diagnosis of the thyroid nodules is very much useful to get knowledge about the death and survival of the cancer cell. This research work has been done on the realistic datasets. The samples of Thyroid nodules tissue were retrieved from 80 patients, who ware recovering from treatment in the form of surgery. This research work has been focused on the thyroid carcinoma. The Bioinformatics tools have been in this research work for exploring the process of identification and the process of characterization of Galectin 3. In this work, at first the RNA has been extracted from tissue samples of Thyroid nodules. After the extraction, the extracted RNA has been transcript in reverse order. After the extraction, partial gene from Galectin 3 has been amplified in non-specific manner and using the adapter module as the primers.

After this process, all of the peR products have been organized sequentially and have been analysed. This process has produced results that contains couples of fragments of exogenous DNA. In another work (Shen, 2020), Bioinformatics has been applied to create biomarkers for thyroid cancer. The mechanism of molecules and the marker of the genetic structure for the thyroid carcinoma is not very clear to the researchers. Therefore, lot of research works have been done to gather knowledge on the genetic structure of the thyroid cancer. This research work has been done on four expression profiles namely GSE3467, GSE33630, GSE3678, and GSE53157. These experimental data have been collected from the 'Gene Expression Omnibus database (GEO)'. This database contains one hundred and sixty-four tissue samples in total in which sixty-four samples are normal tissue sample of thyroid gland and 100 samples are cancer tissue samples of thyroid gland. In this research work, a method named Robust Rank Aggreg (RRA) has been applied to create Differentially Expressed Genes (DEGs). On the created Differentially Expressed Genes (DEGs) four operations have been done sequentially. The four applied operations are – functional annotation using Gene

Ontology (GO), Pathway Analysis, Protein Protein Interaction (PPI) analysis, and Survival Analysis.

Finally, the small molecule candidates containing the genetic information about the thyroid cancer cell has been identified by using CMap. This research work, results a better knowledge on the genetic structure of the thyroid tissues, which contains the cancer cells. J. Tang et al in their work (Tang, 2018), have used the Bioinformatics to understand the genetically structure changes responsible for thyroid cancer. The Thyroid cancer comes under the category of the endocrine malignancies. According to the authors of the paper, huge numbers of microRNAs with mRNAs have been affected in the malicious Thyroid tissues and these abnormalities have been proved with sufficient reasons. In general, in Tumorigenesis the microRNAs and mRNAs have very important role. In this research work, seventy-two microRNAs and one thousand seven hundred sixty-six mRNAs have been identified, which have been expressed differentially between both of the tissues of Normal Thyroid and Cancer Thyroid.

The huge number of microRNAs and microRNAs have been evaluated for the prognostic values of the molecules by applying 'Kaplan Meier Survival curves' using 'log rank test'. In the evaluation process, seven microRNAs namely 'miR-146b', 'miR-184', 'miR-767', 'miR-6730', 'miR-6860', 'miR-196a-2' and 'miR-509-3'have been associated in the final survival. From the seven microRNAs, three microRNAs have been linked with another six mRNAs, which are expressed differentially. The microRNA 'miR-767' has been predicted to the target mRNAs 'COL10A1', 'PLAG1', and 'PPP1R1C'. The microRNA 'miR-146b' has been predicted to the target mRNA 'MMP16', and the microRNA 'miR-196a-2' has been predicted to the target mRNA 'SYT9'. The highest rated 10 hub genes namely 'NPY', 'NMU', 'KNG1', 'LPAR5', 'CCR3', 'SST', 'PPY', 'GABBR2', 'ADCY8', and 'SAA1'have been screened out for the identification of the key genes from the protein protein interaction (PPI) network. The gene named LPAR5 has been associated with the final survival. According to the multivariate analysis, 'miR-184', 'miR-146b', 'miR-509-3', and 'LPAR5' have been found as the independent risk factor for the process of prognosis. The research work has identified a chain of microRNAs and mRNAs that were prognostic and responsible for thyroid cancer. Therefore, these microRNAs and mRNAs can be identified for the treatment of thyroid cancer. The work described in the paper (Liu, 2020) has been done on the 'Papillary thyroid carcinoma (PTC)', which is very frequent thyroid cancer found in human beings.

In the recent years, number of cases of the 'Papillary thyroid carcinoma (PTC)' increases very rapidly. The molecular structure of the gene responsible for the 'Papillary thyroid carcinoma (PTC)' has not been identified properly and there has a major chance of misdiagnosis of the gene molecules. The discussed research work has been aimed to understand the molecular mechanism responsible for 'Papillary thyroid carcinoma (PTC)'. This work has identified key biomarkers for the prognosis. In this work, 'Differentially Expressed Genes (DEGs)' have been explored the normal thyroid tissue and the Papillary thyroid carcinoma (PTC) tissue using the Integrated Analysis. The functionalities and the pathways of the 'Differentially Expressed Genes (DEGs)' have been investigated by three processes namely 'Gene Ontology', 'Pathway', and 'Protein Protein Interaction (PPI)' network analysis. The receiver operating characteristic (ROC) curve has been used for

predicting the accuracy of the 'Differentially Expressed Genes (DEGs)'. Four datasets namely 'GSE33630', 'GSE27155', 'GSE3467', and 'GSE3678' from the 'Gene Expression Omnibus' database have been used for the research work, which contains total 153 'Differentially Expressed Genes (DEGs)' including sixty-six up regulated and eighty-seven down regulated. R. Fan et al have done a research work on the 'Papillary thyroid carcinoma (PTC)', which has been described in the paper (Fan, 2021).

In this work, two online software has been used namely GEO2R and Venn. This two software have been used for the screening of genes, which had been expressed differentially. The Hub genes have been screened using STRING and Cytoscape at first. After that, another screening process of the hub genes has been done by using and FEGG enrichment analysis and Gene Ontology. At the last, survival analysis with expression validation have been performed with the help of the online software namely UALCAN and immune histo chemistry respectively. The research work has identified total number of three hundred and thirty-four Differentially Expressed Genes (DEGs), which ware consistent and the identified Differentially Expressed Genes (DEGs) has contained one hundred and thirty-six up regulated with one hundred and ninety-eight down regulated genes. According to the Gene Ontology enrichment analysis done in this work has suggested the fact that the Differentially Expressed Genes (DEGs) have been enriched in the carcinoma related functionalities and pathways. In this work, Protein Protein Interaction network has been visualized in which seventeen samples of up regulated and thirteen samples of down regulated Differentially Expressed Genes (DEGs) have been selected. The 'Gene Expression Profiling Interactive Analysis Tool (GEPIA)' and 'UALCAN' tool has been used for the expression verification and overall survival analysis respectively has suggested the fact that three hub genes namely 'LPAR5', 'TFPI', and 'ENTPD1' have been responsible for the 'Papillary thyroid carcinoma (PTC)'.

Research Methodology

The process of classifying the genetically data of the thyroid patients has been done in this work using some of the machine learning based classification tool. The classification algorithm is responsible for classifying some samples among several classes. The classification process has mainly two parts – training and testing. In the training process, computational device tries to learn the internal structure of the data. At the time of training, the data with the corresponding class has been feed into the classification model. From the test data, classification algorithm learns about the data and corresponding class. Different classification algorithms use different strategy for learning. The classification model is working on the strategy learned at the time of training process. At the time of testing, the classification task has been done on the learning strategy taken at the time of training process. In this work, we have used three classification algorithms namely Random Forest (RF), k Nearest Neighbour (kNN), and Support Vector Machine (SVM). The three classification algorithms are working on different strategies. Therefore, these classification algorithms have been applied separately on the thyroid patient data for comparing the classification task. Now, the three-classification algorithm have been discussed thoroughly in the following part.

Random Forest (RF)

Random Forest is the classification algorithm, which is created on the concept of decision trees. In this algorithm, couples of decision trees have been grouped to create the random forest. Each of the decision trees produces some decisions and at the final stage, the majority of the decisions has been taken as the final decision. The classification task has been done in very efficient manner if the decisions taken by the decision trees are not very much correlated. The researchers can choose the Random Forest (RF) algorithm for the classification task because this algorithm is the fastest algorithm to train the classification model. This algorithm needs the smallest time for training. This algorithm can handle large dataset as well. In the Bioinformatics based research, large dataset has to be handled. Therefore, Random Forest (RF) becomes very much useful. At the same time, the random Forest (RF) algorithm produce high accuracy in the classification task even it is applying on the large dataset. The Random Forest (RF) algorithm is fault tolerant as well. The algorithm works fine whether a large amount of data from one class has been removed. The Random Forest (RF) algorithm has been applied in two parts. At the first part, n number of random decision trees have been chosen. At the second stage, decisions trees have been used for making the decisions. The final decision has been taken by the process of majority voting from the decisions made by the participating decision trees.

k Nearest Neighbour (kNN)

The k Nearest Neighbour is one of the machines learning based algorithm used for the classification task. The k Nearest Neighbour (kNN) algorithm comes under the category of supervised machine learning. The supervised machine learning is the process where the classifier model has been trained by external matters. That means the supervised classifier has been trained with labelled data. The k Nearest Neighbour (kNN) is the classification algorithm, which has been trained with labelled data. Therefore, it is the supervised algorithm. The kNN is the simplest and easy to coding classification algorithm. Although, k Nearest Neighbour algorithm has been used for classification and regression task; k Nearest Neighbour has been used prominently for the classification task. The processing of the k Nearest Neighbour algorithm is bit different from the other classification algorithms. The k Nearest Neighbour algorithm has not done any calculations at the time of training process. At the training process, the k Nearest Neighbour algorithm only gathers training data with their corresponding class. At the time of testing, the k Nearest Neighbour algorithm starts computing. When a test sample comes to the algorithm, then it starts working and tries to match the test sample with the training sample and adjust the test sample to the class with maximum similarity. For these special characteristics of the k Nearest Neighbour algorithm, this algorithm is known for Lazy Learning algorithm. The main disadvantage of the k Nearest Neighbour algorithm is that in this algorithm the initial k neighbours have to be chosen at random. However, the success of the k Nearest Neighbour algorithm is depends on the initial k neighbours.

Support Vector Machine (SVM)

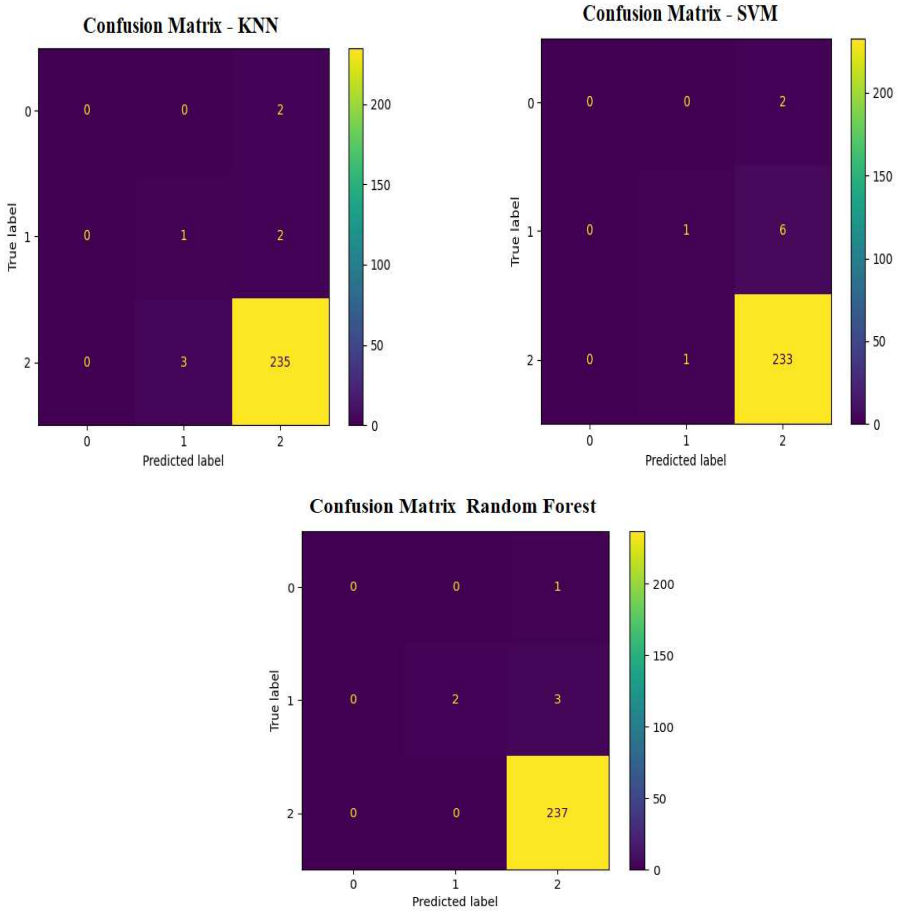
The support Vector Machine is a supervised machine learning algorithm mainly used for classification task. Although it can be used for regression task also. According to the Support Vector Machine algorithm, the observations have been plotted on n dimensional space. The classes have been separated by the hyper-planes drawn on the n dimensional space. The Support Vector Machine algorithm has chosen the extreme locations from the input vectors to draw the hyperplane. The extreme locations are known as the support vectors. The algorithm has been named after the support vectors. Choosing the support vectors are the main mechanism in the algorithm. By using the support vectors, decision boundary can be created. According to the hyperplane divided the classes, the Support Vector Machine (SVM) has been divided in to two types – liner Support Vector Machine (SVM) and nonlinear Support Vector Machine (SVM). In case of the linear Support Vector Machine (SVM), the hyperplanes drawn are linear in character. However, in case of the nonlinear Support Vector Machine (SVM), the hyperplane drawn are nonlinear in characteristic. The complexity of the classification problem is depending on the characteristics of the hyperplane drawn. When the classes are separable enough, then the hyperplane must be linear. When the classes are very much related, then nonlinear hyperplanes should be drawn. The shape of the hyperplane is dependent on the number of features used in the training process. If the feature size is two then the hyperplane must be a straight line. If the feature number increases then the hyperplane becomes n dimensional figure. The Support Vector Machine (SVM) is the most efficient classification algorithm and can solve critical problems.

Result and Analysis

With the implementation of Thyroid Disease Data Set in python environment this proposed work is set up. This dataset is collected from UCI Machine Learning Repository. For a classification-based machine learning work, several modelling approaches needs to be implemented and thus, in this scenario as well three machine learning approaches will be presented. These approaches are random forest, SVM and KNN. These approaches are utilized and these all are individually making prediction to find out thyroid disease. For the successful completion of the proposed task, first of all the data was loaded and observed closely to understand the nature of data, class information as well as the size of data. To make the efficient learning by the said models the information needs to get cleaned and converted into numerical digits. Then it has been divided into two parts: training part and testing part. First the implemented models are one by one called, get trained with the training dataset and then these are all tested with test dataset. Once they are gone over the tested dataset, they come up with their individual classification report. Form these reports made by the three machine earning approaches, we made out comparison so that a best fitted model is determined by the performance in terms of prediction. The accuracy scores of the individual implemented classification techniques are following:

- The random forest approach resulted 98% of accuracy.
- The SVM approach resulted 96% of accuracy.
- The KNN approach resulted 97% of accuracy.

To understand the classification reports closely and precisely the graphical representation of the classification report ha been represented by the confusion matrix of all of the three implemented models. Confusion matrices are consisting of four matrices – true positive, true negative, false positive and false negative. The model wise confusion matrices are as follows:



Discussion

For the classification in this proposed model, popular supervised machine learning approaches like random forest, SVM and KNN has been used to predict the dataset of Thyroid Disease. It has been observed that all of the implemented models are scoring their best accuracy score and these are all very close to others reports as well. But comparing all of them we found that Random Forest model has performed a bit better than the other implemented model in this said dataset. So, it can be said that for this suggested model this Random Forest approach is the best fitted approach. In the future we may increase the size of the dataset and check the created model with benchmark testing to validate the model in every aspect.

Conclusion

At the completion of the proposed task, it can be said that a successful implementation of the classification of the thyroid disease has been performed here and by comparing them we represented an approach where a comparative study has been performed. Later on, there is a huge scope to continue this research work in various way. This sort of works is helpful to determining the thyroid disease and save a life from affected of this disease. Hence, we also represented with the importance of bioinformatics and its utilities and various industries and medical industry is one of them.

References

Irizarry, Lisandro. "Thyroid hormone toxicity." *Medscape. WedMD LLC. Retrieved* 2 (2014).

Kim, Ha-Young, and Subburaman Mohan. "Role and mechanisms of actions of thyroid hormone on the skeletal development." *Bone research* 1.1 (2013): 146-161.

Zheng, Long, et al. "Bioinformatics analysis of key genes and pathways in Hashimoto thyroiditis tissues." *Bioscience reports* 40.7 (2020).

Zhi-long, Cheng, et al. "Application of a bioinformatics method on detecting the biomarkers to predict thyroid nodules diagnosis." *2012 IEEE Symposium on Electrical & Electronics Engineering (EEESYM)*. IEEE, 2012.

Shen, Yujie, et al. "Identification of potential biomarkers for thyroid cancer using bioinformatics strategy: a study based on GEO datasets." *BioMed Research International* 2020 (2020).

Tang, Jianing, et al. "Bioinformatic analysis and identification of potential prognostic microRNAs and mRNAs in thyroid cancer." *PeerJ* 6 (2018): e4674.

Liu, Y., Gao, S., Jin, Y., Yang, Y., Tai, J., Wang, S., ... & Guo, Y. (2020). Bioinformatics analysis to screen key genes in papillary thyroid carcinoma. *Oncology letters*, *19*(1), 195-204.

Fan, Rong, et al. "Integrated bioinformatics analysis and screening of hub genes in papillary thyroid carcinoma." *Plos one* 16.6 (2021): e0251962.

Suryasa, I. W., Rodríguez-Gámez, M., & Koldoris, T. (2022). Post-pandemic health and its sustainability: Educational situation. *International Journal of Health Sciences*, *6*(1), i-v. https://doi.org/10.53730/ijhs.v6n1.5949

Suryasa, I.W., Sudipa, I.N., Puspani, I.A.M., Netra, I.M. (2019). Translation procedure of happy emotion of english into indonesian in kṛṣṇa text. *Journal of Language Teaching and Research, 10*(4), 738–746