Implementation Of K-Nearest Neighbor - Certainty Factor For Expert System Detection Of Idiopathic Thrombocytopenic Purpura

Eva Y Puspaningrum*, Budi Nugroho, Dwi Putri Safira
Informatics Faculty of Computer Science, Universitas Pembangunan Nasional “Veteran” Jawa Timur, Indonesia.

Abstract. Idiopathic Thrombocytopenic Purpura (ITP) is an autoimmune disorder. ITP can occur in children and adults. This disease can be fatal because the platelet count is low due to the destruction of excessive platelets so that it can interfere with vital organs and bleeding occurs. The lack of knowledge of ordinary people about ITP disease, so many people assume that bruises and nosebleeds on the body are caused by fatigue. For that, we need a system that can imitate the expertise of an expert in diagnosing this disease based on the symptoms felt. The method used to support the expert system is the K-Nearest Neighbor and Certainty Factor methods which are a combination of 2 methods, where the classification results from the K-Nearest Neighbor method will be given a certainty value by the Certainty Factor method so as to produce a prediction. The results of combining the two methods can produce certainty in the diagnosis. Based on the test results using 3 test scenarios using parameter values k=3, k=5, k=7 and the results obtained the highest accuracy value with parameter value k=7 obtained an accuracy rate of 90.9%.

Keywords: Idiopathic Thrombocytopenic Purpura, Expert System, KNN, Certainty Factor

1 Introduction
Idiopathic Thrombocytopenic Purpura (ITP) is an autoimmune disorder in which autoantibodies are directed against the patient's own platelets causing destruction and bleeding of platelets [1]. If the damage is severe it can result in death due to blood loss or bleeding in vital organs. ITP can occur in children and adults. ITP does not appear to be related to race, lifestyle, climate or environmental factors [2]. ITP can last for weeks or months and can turn into a chronic condition. As many as 2-5% of people with ITP are associated with a high risk of bleeding, such as cerebral hemorrhage and often require restriction of physical activity [3]. The lack of knowledge of people about ITP and still think that bruises and nosebleeds on their bodies are just because of fatigue. This problem could be a symptom of ITP. Because the symptoms that are considered only bruises make someone reasons why people are reluctant to check themselves, so a system that is designed that has the ability of a doctor is needed in diagnosing a disease. As technology develops, it can affect all activities in various fields of life. One of them is in the health sector. With the expert system can help diagnose the disease based on the symptoms felt by the patient, which then produces information about the disease more quickly. In the field of informatics, there are several methods used to help optimize expert system applications so that they can provide high accuracy in early diagnosis. One method of object classification is K-Nearest Neighbor (KNN). This classification is based on the data that is closest to the object. KNN can also be defined as a classifier that is used to classify data based on the comparison of the K values of the nearest neighbors, the K parameter in KNN has a major influence on the final prediction results generated [4]. The method used to support the expert system is the K-Nearest Neighbor and Certainty Factor methods which are a combination of 2 methods, where the classification results from the KNN method will be given a certainty value by the Certainty Factor method so as to produce a prediction. The Certainty Factor method is embedded in this system. The CF method is used because the performance of the system can run according to functional requirements and has a high percentage of accuracy [5]. In addition, the CF method can describe the level of expert confidence in the problems at hand.

2 Literature Review

2.1 Idiopathic Thrombocytopenic Purpura (ITP)
Idiopathic thrombocytopenic purpura (ITP) is a chronic autoimmune disease accompanied by bleeding. an autoimmune disorder that attacks platelets causing destruction and bleeding of platelets. Decreased platelet
count due to increased destruction is the main cause of disease [6]. This disease is the most common autoimmune disorder involving blood cells. ITP causes a decrease in the number and function of platelets which causes bleeding is the most important complication of ITP. Bleeding is mostly seen on the skin and mucosa. and the most important complications of the disease include anemia and extensive bleeding [7]. In the case of a significant decrease in platelet count, the risk of profuse bleeding increases with the development of severe anemia. The chronic form of ITP is more likely to occur in adults. The disease often develops without an explicit association with a previous disease.

2.2 K-Nearest Neighbor

K-Nearest Neighbor (K-NN) is an object classification method based on data from the closest distance value learning. KNN used to overcome the identification of measured values both qualitatively and quantitatively [8]. KNN is a supervised learning algorithm where the results of a new query instance are classified based on most of the categories in the KNN algorithm. Where is the class that appears the most that will be the class that results from the classification? The class with the most of the nearest k data will be selected as the predicted class for the new data. In general, the value of k uses an odd number so that there are no equal distances in the classification process. luckily calculates distance using Euclidean distance.

2.3 Certainty Factor

The certainty factor (CF) is one of the techniques used to overcome uncertainty in decision making. The certainty factor (CF) can occur under various conditions. Among the conditions that occur there are several different rules with the same consequences. In this case we have to combine CF. The total value of the existing conditions [9]. The certainty factor uses a value to assume the degree of confidence that an expert has in a data can be seen in the equation (1).

\[ CF(\text{rule}) = MB(H, E) - MD(H, E) \]  

(1)

CF is a certainty factor, MB is a measure of Belief, MD is a measure of Disbelief, H is Probability and E is Evidence.

3 Methodology

The data used in this study contains data on the symptoms experienced by patients with Idiopathic Thrombocytopenic Purpura (ITP) at the Jombang Regency General Hospital. Data on ITP symptoms and attribute values from experts can be seen in Table 1 and Table 2 is the categories of diagnostic results and solutions provided by experts.

| No  | Code | Symptom                           | Expert Value |
|-----|------|-----------------------------------|--------------|
| 1   | GA01 | Unexplained bruising on body parts | 0.8          |
| 2   | GA02 | Prolonged bleeding due to wound    | 0.4          |
| 3   | GA03 | Bleeding that occurs on the skin of the legs is always purplish red spots | 0.8 |
| 4   | GA04 | Nose always nosebleed             | 0.2          |
| 5   | GA05 | Bleeding in urine or stool         | 0.4          |
| 6   | GA06 | Bleeding gums                      | 0.6          |
| 7   | GA07 | Excessive or prolonged bleeding    | 0.6          |
| 8   | GA08 | Always fatigue                     | 0.4          |

Table 2. Categories of Diagnostic Results and Solutions Provided by Experts

| No  | Class      | Solution                                                                 |
|-----|------------|--------------------------------------------------------------------------|
| 1   | Mild ITP   | Adequate rest, eat regularly that contains to increase platelets.         |
| 2   | Chronic ITP| Must eat foods that can increase platelets, use of steroids (by prescription) used to alter or simulate the effects of hormones, often used to reduce inflation or for tissue repair and growth, blood transfusions are performed. |

The knowledge base contains information about relevant knowledge needed to understand, formulate and solve knowledge base problems What is the core program of the system, where the knowledge base is a representation of an expert's knowledge shown in Table 3.

| No | Rule                                                                 |
|----|----------------------------------------------------------------------|
| 1  | IF always bruises on body parts without cause (G01) Nose always nosebleeds (G04) And always tired (G08) Then Possible Mild ITP |
| 2  | IF always bruising on parts of the body without cause (G01) And Bleeding lasts a long time due to cuts (G02) And Bleeding that occurs on the skin of the legs always purplish red spots (G03) And Bleeding in urine or stool (G05) And Bleeding in the gums (G06) And excessive or prolonged bleeding during menstruation (G07) Then Possible Chronic ITP |

In the K-Nearest Neighbor - Certainty Factor method, the weight value of each answer selected by the user comes from the term table of user interpretation values. In the classification process there are several stages in pre-processing the data, to retrieve the relevant data used as research and convert the categorical data into numeric. In collecting data using the user's answer categories such as “Very Confident”, “Confident”,...
“Quite Sure”, “Less Sure”, and “No”. and then it was changed to numeric namely "Very Confident to be 1", "Confident to be 0.8", "Quite sure to be 0.6", "Less sure to be 0.4", and "No to be 0". In this research first step is pre-processing data, after that the data is divided into training data and test data which will be calculated using the KNN method to determine the classification. After the classification process is carried out, the certainty value of the classification will be calculated using the certainty factor method. The next step is to calculate the certainty value using CF.

4 Results

From the scenario, it will be tested on 3 different k values. In the k parameter variation test, it is carried out to find out whether it can affect the results of the ITP diagnostic system capability test scores that have been made. Testing is done by comparing the results of the system diagnosis with the results of expert diagnoses or the original condition in accordance with the parameter values of k=3, k=5, and k=7. In the implementation of testing the accuracy of the system using 12 test data which is random data. The results with parameter k = 3 are presented in Table 4.

| No. | Name | Expert Result | System Result |
|-----|------|---------------|---------------|
| 1   | A    | Chronic ITP   | Chronic ITP   |
| 2   | B    | Chronic ITP   | Chronic ITP   |
| 3   | C    | Chronic ITP   | Chronic ITP   |
| 4   | D    | Mild ITP      | Mild ITP      |
| 5   | E    | Mild ITP      | Chronic ITP   |
| 6   | F    | Chronic ITP   | Chronic ITP   |
| 7   | G    | Chronic ITP   | Mild ITP      |
| 8   | H    | Chronic ITP   | Chronic ITP   |
| 9   | I    | Mild ITP      | Mild ITP      |
| 10  | J    | Mild ITP      | Mild ITP      |
| 11  | K    | Mild ITP      | Mild ITP      |

From the test results in Table 4 it produces 5 True Positives, which means that in the diagnostic test they fall into the category of Chronic ITP and the system has a true value that they suffer from Chronic ITP, other results are 4 True Negatives, which means they are in the diagnostic test into the category of Mild ITP, and in fact they actually suffer from Mild ITP. Meanwhile, there are test results which state that they are actually mild ITP but the results of the Chronic ITP test are thus declared 1 False Negative. And there are test results stating that they are actually Chronic ITP but the results of the Mild ITP test are thus declared 1 False Positive. From the calculation of the test data produces an accuracy value of 81.81% by using the parameter value k=3.

| No. | Name | Expert Result | System Result |
|-----|------|---------------|---------------|
| 1   | A    | Chronic ITP   | Chronic ITP   |
| 2   | B    | Chronic ITP   | Chronic ITP   |
| 3   | C    | Chronic ITP   | Chronic ITP   |
| 4   | D    | Mild ITP      | Mild ITP      |
| 5   | E    | Mild ITP      | Chronic ITP   |
| 6   | F    | Chronic ITP   | Chronic ITP   |
| 7   | G    | Chronic ITP   | Mild ITP      |
| 8   | H    | Chronic ITP   | Chronic ITP   |
| 9   | I    | Mild ITP      | Mild ITP      |
| 10  | J    | Mild ITP      | Mild ITP      |
| 11  | K    | Mild ITP      | Mild ITP      |

From the results of the accuracy test for scenario 2 presented in Table 5 using the parameter value k = 5 which produces 6 True Positives, which means they are in the diagnostic test into the Chronic ITP category and the system has a true value that they suffer from Chronic ITP, other results are 3 True Negative, which means that in the diagnostic test they fall into the category of Chronic ITP and in fact they actually suffer from Mild ITP. Meanwhile, there are test results that state that they are actually mild ITP but the results of the Chronic ITP test are thus declared 2 False Positive. From the calculation with the parameter value k=5 the accuracy value is 81.81%.

| No. | Name | Expert Result | System Result |
|-----|------|---------------|---------------|
| 1   | A    | Chronic ITP   | Chronic ITP   |
| 2   | B    | Chronic ITP   | Chronic ITP   |
| 3   | C    | Chronic ITP   | Chronic ITP   |
| 4   | D    | Mild ITP      | Mild ITP      |
| 5   | E    | Mild ITP      | Chronic ITP   |
| 6   | F    | Chronic ITP   | Chronic ITP   |
| 7   | G    | Chronic ITP   | Mild ITP      |
| 8   | H    | Chronic ITP   | Chronic ITP   |
| 9   | I    | Mild ITP      | Mild ITP      |
| 10  | J    | Mild ITP      | Mild ITP      |
| 11  | K    | Mild ITP      | Mild ITP      |

From the results of the scenario 3 test presented in Table 6 using the parameter value k=7 and producing 6 True Positive which means those in the diagnostic test fall into the Chronic ITP category and the system has a correct value that they suffer from Chronic ITP, the other results are 3 True Negatives, which means they in the diagnostic test fall into the Chronic ITP category and in fact they actually suffer from Mild ITP. Meanwhile, there are test results stating that they are actually mild ITP but the results of the Chronic ITP test are thus declared 1 False Positive. From the calculation of the test data produces an accuracy value of 90.9% by using the parameter value k=7.

The results of combining the two methods can produce certainty in the diagnosis. Based on the test results using 3 parameter test scenarios the highest accuracy value with parameter value k=7. The results of the trial evaluation can be seen in Table 7.
Table 7. System Performance Evaluation Results

| No | Parameter | Accuracy | Precision | Recall | Specificity |
|----|-----------|----------|-----------|--------|-------------|
| 1  | K=3       | 81.81%   | 83.33%    | 83.33% | 80%         |
| 2  | K=5       | 81.81%   | 100%      | 75%    | 60%         |
| 3  | K=7       | 90, 9%   | 100%      | 85.7%  | 80%         |

5 Conclusions

In the expert system system for detection of ITP disease using the KNN – CF method produces a relatively good level of accuracy. The KNN method will classify the detection results by calculating the shortest distance after which the classification results will be assigned a certainty value using the CF method. From several tests with the value of the parameter variation, the highest accuracy rate is 90.9% at the variation of k=7.

References

[1] Chunyan Hea, Lian Zhaoa, Yumei Niea, Rong Yana, Kangxi Zhoua, Xiaodong Lia, Birong Zhoua, Kesheng Daia, 2020 A novel method to detect autoantibodies against platelets in patients with immune thrombocytopenia, Clin. Chim. Acta vol. 511 September pp. 90–93 doi: 10.1016/j.cca.2020.09.029.

[2] Naima Al-Mulla, Abdulbari Bener, Aliaa Amer, Mohammed Abu Laban, Idiopathic thrombocytopenic purpura in childhood: A population-based study in Qatar, J. Pediatr. (Rio. J) vol. 85 no. 3 pp. 269–272, (2009), doi: 10.2223/JPED.1851.

[3] Chae Young Kim, Eun Hye Lee and Hoi Soo Yoon, 2016 High remission rate of chronic immune thrombocytopenia in children: Result of 20-year follow-up, Yonsei Med. J vol. 57 no 1, pp. 127–131 doi: 10.3349/ymj.(2016).57.1.127.

[4] Han J, Pei J, and Kamber M. 2011 Data mining: concepts and techniques. Elsevier: Amsterdam

[5] Aryu H Aji, M. Tanzil F, and Agus Wahyu W, Sistem Pakar Diagnosa Penyakit Ibu Hamil Menggunakan Metode Certainty Factor (CF), J. Pengemb. Teknol. Inf. dan Ilmu Komput., vol 3 no 5, pp 2127–2134, (2018).

[6] Mantadakis E, Farmaki E, Buchanan GR. Thrombocytopenic purpura after measles-mumps-rubella vaccination: a systematic review of the literature and guidance for management. J Pediatr. (2010);156 (4):623–8.

[7] Ulugbek Y. E, Gaybulla I.M, Aminjon M.K, Davron R.Y, Fatima A.A, Shirinkhon G. A Treatment of Idiopathic Thrombocytopenic Purpura: Prevention of Hormonal Complications in The Stomach and Duodenum, Central Asian Journal of Medicine vol. 2020, no. 2, (2020).

[8] F. H. Tawarai, F. Fauziah, and A. Andrianingsih, Web-Based Rice Disease Diagnosis Expert System Using Fuzzy Tsukamoto Method and K-Nearest Neighbor Algorithm, J. Comput. Networks, Archit. High Perform. Comput., vol. 3, no. 2, pp. 153–160, (2021), doi: 10.47709/cnahpc.v3i2.980.

[9] A. Khozaimi, Mobile Expert System for Dengue Fever Based on Certainty Factor Method, J. Phys. Conf. Ser., vol. 1569, no. 2. (2020) doi: 10.1088/1742-6596/1569/2/022070.