ISSN No:-2456-2165

Optimization of Convolutional Neural Network Architecture with Genetic Algorithm for Feature Extraction Prediction of Student Graduation

Abi Mestu Yudansha[1]; M Arief Soeleman[2]; Affandy
Magister Teknik Informatika, Universitas Dian Nuswantoro Semarang, Indonesia

Abstract:- Higher education institutions such as universities play a central role in the steps where comprehensive research and development activities take place within a highly competitive environment. The academic achievements of students become a crucial element within the structure of these higher education institutions. This is because one of the key indicators of university quality is an exceptional track record of academic achievements. Universitas Dian Nuswantoro (UDINUS), a private educational institution with an A accreditation rating, is located in Semarang, Indonesia. One of the faculties that holds a significant role at UDINUS is the Faculty of Computer Science, which stands out with the highest number of students, particularly in the Bachelor's program of Computer Science (S1), which records a comprehensive and outstanding student count compared to other study programs. Therefore, it is appropriate to focus research on the data regarding the graduation rates of students from the Computer Science S1 program. In this study, the author applies Data Mining, a method involving manipulation of large-scale data. The primary mission of this research is to address the question of how the implementation of Deep Learning using an optimized Convolutional Neural Network (CNN) through Genetic Algorithm can be utilized to predict student graduation. Consequently, the outcomes can serve as references to expedite student graduation. The study demonstrates that the feature extraction values using CNN and the hyperparameters using Genetic Algorithm show an overall increase in accuracy when using K-Nearest Neighbor (K-NN) for all values of n: 3, 4, 5, and 6 (Proven that feature extraction from tabular data represented as images and processed with the CNN algorithm using the most suitable parameters is successful)

Keywords:- *Component Udinus, Convolutional Neural Network (CNN), Genetic Algorithm, K- Nearest Neighbor (KNN).*

I. INTRODUCTION

Higher education institutions such as universities are at the heart of the measures where extensive research and development is carried out in a competitive environment. Universities usually need knowledge gathered from past and current data sets that can be used to represent and convey information to university admins to monitor conditions and take steps to resolve problems [1].

One of parameters which indicates superior quality of a university is a history of academic achievements. Student achievement can be assessed through an assessment of the learning process and curriculum applied. Most research has taken the focus point that graduates are the most important indicator of student success. Problems related to graduation rates are something that is often experienced by higher education institutions. Predicting graduation rates still contains challenges, so management to avoid a decrease in graduation rates is difficult.

There are various factors that have an influence on the quality of a university. One of them is the success rate of students in completing their studies on time. The increasing number of students who can complete their studies on time will have a positive impact on program accreditation assessments in higher education. There are many aspects that influence graduation rates, including inadequate levels of academic ability, course program design, academic achievement, and other factors. Limited access for students from remote areas to reliable internet connections, quota problems, and challenges in understanding in-depth course material are things that also influence, or in other words, dissatisfaction with distance (online) learning.

All of these things will likely have a negative impact on student graduation rates in the future. Dian Nuswantoro University (UDINUS) is a large private academy with A accreditation located in Semarang, Indonesia. Founded in 1990, guided by Mr. Prof. Dr. Ir. Edi Noersasongko, M.Kom. Prof. Edi said, "Our university is growing rapidly and becoming a facilitator of quality higher education. UDINUS is committed to sharing learning for each student and life path" [2]. One of the faculties available and an advantage of UDINUS is the Faculty of Computer Science. This faculty has the largest number of students, especially in the Undergraduate (S1) Informatics Engineering study program, which has a higher diversity and number of

students compared to other study programs. Therefore, it is relevant if data regarding student graduation from the Undergraduate Information Engineering study program is considered a worthy subject for research.

Data mining is a process that involves manipulating data on a large scale. Data mining as steps in uncovering data structures to obtain solutions. These data structures can take various forms, including rules, graphs, networks, trees, and equations, as well as other forms. Using data mining techniques, cases can be analyzed to identify trends, structure, and even future predictions. One of them is classification, which aims to group data based on preexisting patterns. The principle is that data with similar structures tend to be similar.

Machine Learning is an algorithmic approach in which a computer system is programmed to process data, learn from such data, and use the insights gained to make decisions or take actions without needing to be explicitly programmed for each such action. Deep Learning is a part of Machine Learning which is designed to continuously analyze data with a logical structure similar to how humans make decisions [3]. Convolutional Neural Network (CNN) is a Deep Learning method used to detect and recognize objects in digital images. CNN's capabilities are claimed to be the best model for solving Object Detection and Object Recognition problems. In 2012, research on CNNs was able to perform digital image recognition with accuracy that rivals humans on certain datasets [4].

Genetic Algorithm (GA) is part of the Evolutionary Algorithm group, a type of algorithm that adopts the process of natural evolution as its inspiration. The main principle in GA is that individuals who have superior qualities tend to survive, while individuals who have less quality are likely to become extinct. This individual superiority mechanism is assessed through a function known as the fitness function. In GA, this concept of fitness is defined as an assessment of the extent to which a solution is suitable for solving a particular problem. This fitness function produces a fitness value that reflects the level of suitability of the solution to the problem at hand. This fitness value is then used as a guide in selecting individuals who will inherit the next generation or in creating new individuals through operations such as reproduction, crossover and mutation. Thus, the population evolutionarily changes towards an increasingly optimal solution as the iteration progresses.

II. METHODOLOGY

A. Selecting a Template (Heading 2)

Type of research is a quasi-experiment using a dataset of undergraduate graduates at the Faculty of Computer Science, Dian Nuswantoro University. This research aims to implement a deep learning convolutional neural network that can predict student graduation and is expected to be able to provide solutions so that students can graduate on time. The research steps are collecting data, formulating hypotheses or propositions, testing hypotheses, interpreting results and conclusions [5].

The dataset used is the undergraduate study graduation dataset at the Faculty of Computer Science, Dian Nuswantoro University, totaling 4041 records. The graduate data period is between the entry years 2012 to 2018 from 4 study programs. Divided into 2 class categories, namely graduating on time with 2073 records and graduating not on time with 1968 records. The 12 attributes used consist of Study Program, Home City, Age, Marital, Number of scholarships, Number of leave requests, Number of student activities, number of achievements and GPA for semester 1 to semester 4 [6][7].

Table 1: Atribut Dataset

Attribute name	Description
Study Program	4 Study program (11, 12, 14 and 15)
Home city	1 : from Semarang 2 : from other city
Age	1 : age> 0 & age< 12 2 : age>= 12 & age < 26 3 : age >= 26 & age <46 4 : age >= 46
Marital status	1 : married 2 : not married yet
Number of scholarship	Number of scholarship received
Number of leave requests	Number of students have take a leave
Number of student activities	Number of student activities followed
Number of achievement	Number of achievement ever obtained
GPA 1	GPA Semester 1
GPA 2	GPA Semester 2
GPA 3	GPA Semester 3
GPA 4	GPA Semester 4

[a]Sample of a Table Footnote. *(Table footnote)*

Convolution Neural Network or what is usually abbreviated as (CNN) is a development method from Multi Layer Perception or abbreviated as (MLP). CNN has a greater number of dimensions compared to the MLP method. Convolution Neural Networks have input arrays starting from 2 or more dimensions. Then, the second dimension represents every important part such as the front-back corner. Adjustments are made to the Length * Width * Thickness matrix arrangement. Next, feature extraction is carried out using the CNN architecture to obtain the desired data output. Next, from this data, the CNN architecture is optimized for hyper parameters using a genetic algorithm before being classified using K-NN and evaluated to determine shortcomings and future needs.

CNN is a development of the Multi Layer Perceptron (MLP) method. However, CNN has more dimensions when compared to MLP. CNN receives input in the form of an array with a varying number of dimensions, ranging from two dimensions to more. For example, in tomato image processing, the first dimension of the input may represent the shape of the tomato skin, while the other dimensions may represent specific angles on the tomato. Next, the second dimension represents every important part such as the front and back corners. Adjustments are made to the Length * Width * Thickness matrix arrangement (according to the image channel), illustration:

In the case of data (tabular data) which is processed for graduation data, to match the format above, the data representation is carried out as follows: N is the number of 4041 student graduation data with the P attribute (Length) is a grouping of 12 attributes into 3 groups as follows:

- Student Personal Data Attributes: (Age, Marital, Home City and Study Program)
- History Attributes: (Number of Leave Requests, Number of Scholarships, Number of Student Activities and Number of Achievements / Award Certificates)
- Academic Value Attributes: (Achievement Index Semester 1 – _Semester 4: GPA1, GPA2, GPA3, GPA4).

L attribute (Width) is the number of attributes from each group above (a total of 4 attributes from each group), and K (Thickness) has a value of 1 because it only processes 1 group of data. From the conditions above it can be concluded that we have N*P*L*K = 4041 * 3 * 4 * 1, we have obtained a format from the data which was initially structured (tabular) into a format like an image for the Convolution process (CNN).

In Convolutional Neural Networks (CNN), Feature Extraction is performed by applying a series of convolutional layers to the input data. Each convolutional layer consists of a set of filters or kernels that combine with the input data to produce a set of feature maps that highlight different aspects of the input data.

Overall, feature extraction is an important step in using genetic algorithms for prediction. By reducing the dimensionality of input data, it can help improve the efficiency and accuracy of the optimization process, resulting in better predictive models.

Hyperparameters are parameters that are determined before the machine learning model training process begins. This is different from machine learning model parameters which can be discovered through the training process. Hyperparameters influence how machine learning models learn data and how they make decisions. The use of hyperparameters in machine learning is to help determine the performance of a machine learning model. Some hyperparameters that are commonly used in machine learning include the learning rate in the gradient descent algorithm, the number of iterations in the training process, and the number of trees in the random forest algorithm.

Inappropriate hyperparameters can negatively affect the performance of a machine learning model. For example, if the learning rate is too small, the training process will take a long time to achieve good results. On the other hand, if the learning rate is too large, the machine learning model can jump past the optimal solution and experience overshooting. The hyperparameters on the blue layer are optimized, while the gray layer has fixed hyperparameters.

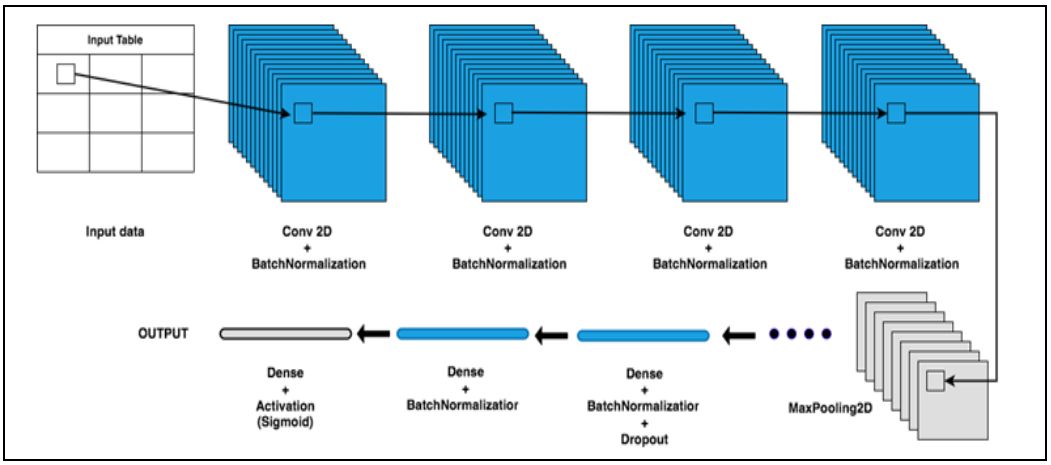

Fig. 1: Structure of CNN

III. RESULT AND DISCUSSION

A. *Analyse of Prepossessing*

The dataset used is a dataset of graduates from the A11 (Informatics Engineering– Bachelor's degree) study program at the Faculty of Computer Science, Dian Nuswantoro University, totaling 2293 records. The graduate data period is between 2012 and 2018. Divided into 2 class categories, namely graduating on time with 1257 records and graduating not on time with 1036 records.

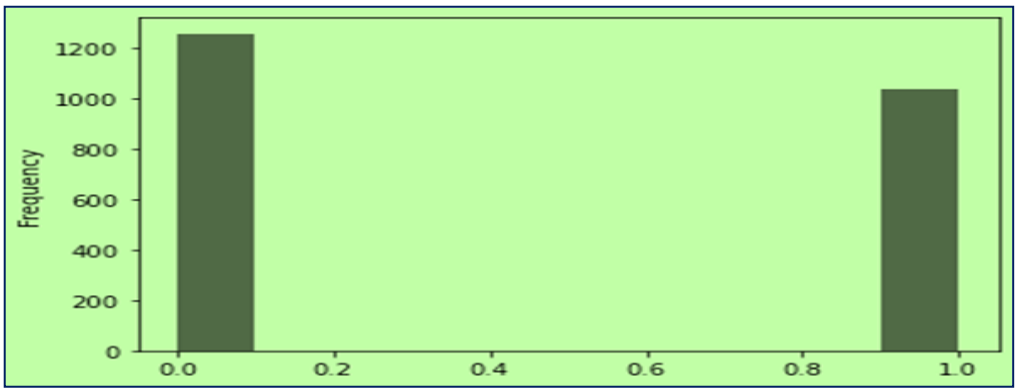

Fig. 2: Graphic Label of Data Set

B. *Analyse of Deep Learning Method CNN*

➢ *Research Step*

CNN format data representation Adjustments for Length * Width * Thickness matrix arrangement (according to image channel), illustration:

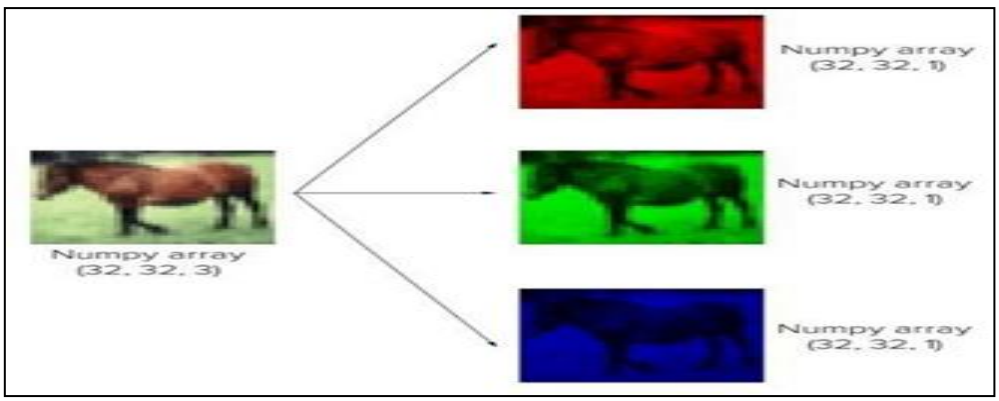

Fig. 3: Research Step

C. *Hyperparameter and Architecture of CNN*

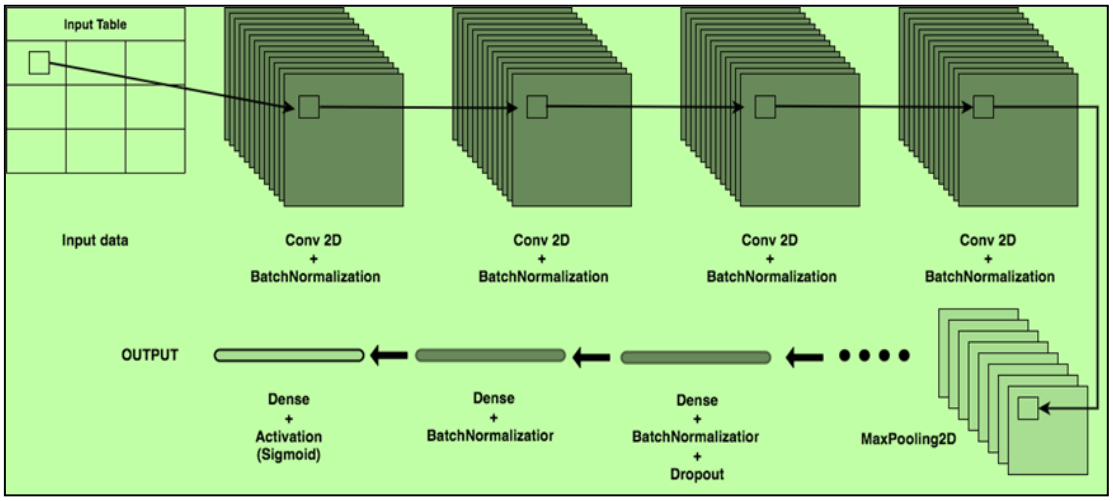

Fig. 4: Hyperparameter and Architecture of CNN

Hyper parameters and architecture in a Convolutional Neural Network (CNN) are elements that influence how a CNN model behaves and learns from data. Choosing the right hyper parameters and designing an appropriate architecture is essential to achieve good performance in a given task CNN Structure. Hyper parameters on the blue layer are optimized, while the gray layer has fixed hyper parameters.

D. Setup Analyse GA

Genetic Algorithm (GA) is an optimization method inspired by the principles of natural evolution. GA is used to find solutions that are close to optimal in complex or multidimensional problems. Genetic Algorithm is used to obtain optimal CNN parameter values. These parameters are the number of hidden layers and time step. Automatically, hidden layers and time steps will be determined by the final results of the Genetic Algorithm. The value of the hidden layer and time step will affect the accuracy results and computing time [22]. Therefore, by adding the Genetic Algorithm, minimum RMSE results and fast computing time will be obtained.

E. Operator on GA Technique

➢ *Selection Using Roulette Wheel*

The roulette wheel selection method is one of the techniques commonly used in Genetic Algorithms to select individual parents from a population when carrying out reproduction. This method selects individuals with a probability proportional to their fitness value. The higher the fitness value, the greater the chance of the individual being selected as a parent. Here's a simple implementation of the roulette wheel selection method using Python.

```
def roulette_wheel_selection(population, fitness_values): total_fitness = sum(fitness_values)
selection_probs = [fitness / total_fitness for fitness in fitness_values]
```

➢ *Cross Over Uses Arithmetic Cross-Over with Definition*

Arithmetic crossover is one of the crossover methods used in Genetic Algorithms to produce individual offspring (children) from two individual parents. This method is often used to combine the characteristics of parents and produce offspring that are a linear combination of gene values in both parents. The steps in arithmetic crossover are as follows:

- Choose parents: Select two individual-parents from the population.
- Choose the r value: Choose an r value between 0 and 1. This r value will control how much influence each parent has on the offspring
- Choose crossover at index gen 1 and 2.

➢ *Mutation Uses Random Mutation where the Mutated Gene is Randomized*

The mutation method using random mutation is a technique commonly used in Genetic Algorithms to introduce genetic variation in populations. In this method, several genes in an individual are determined to be mutated, and the values of these genes are re-randomized with random values.

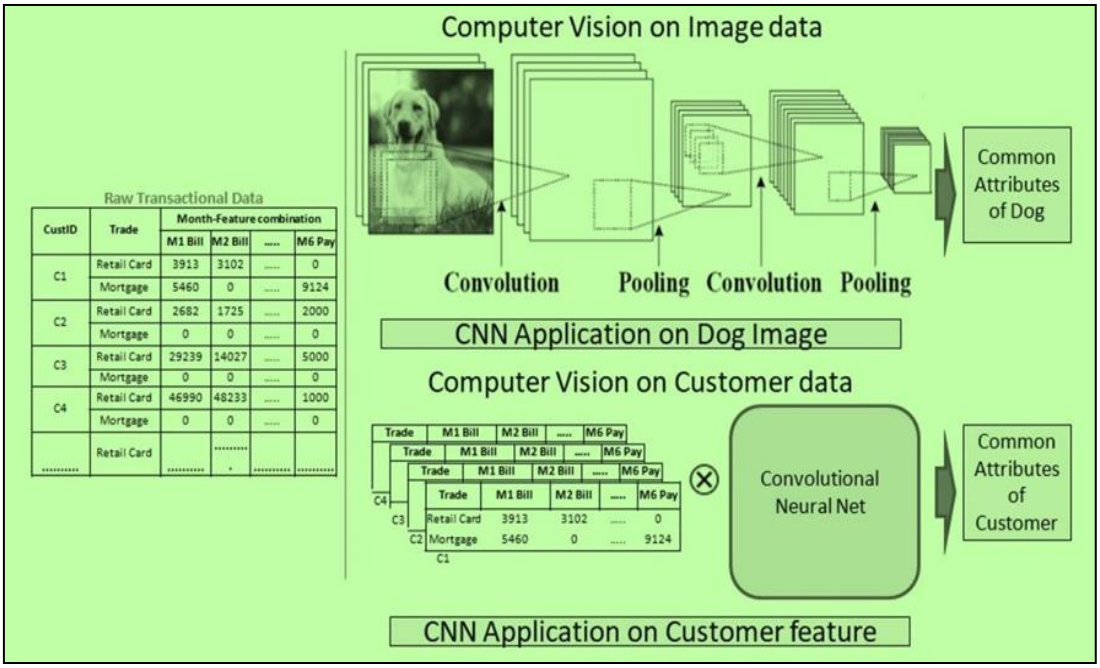

Fig. 4: Extraction Fixture

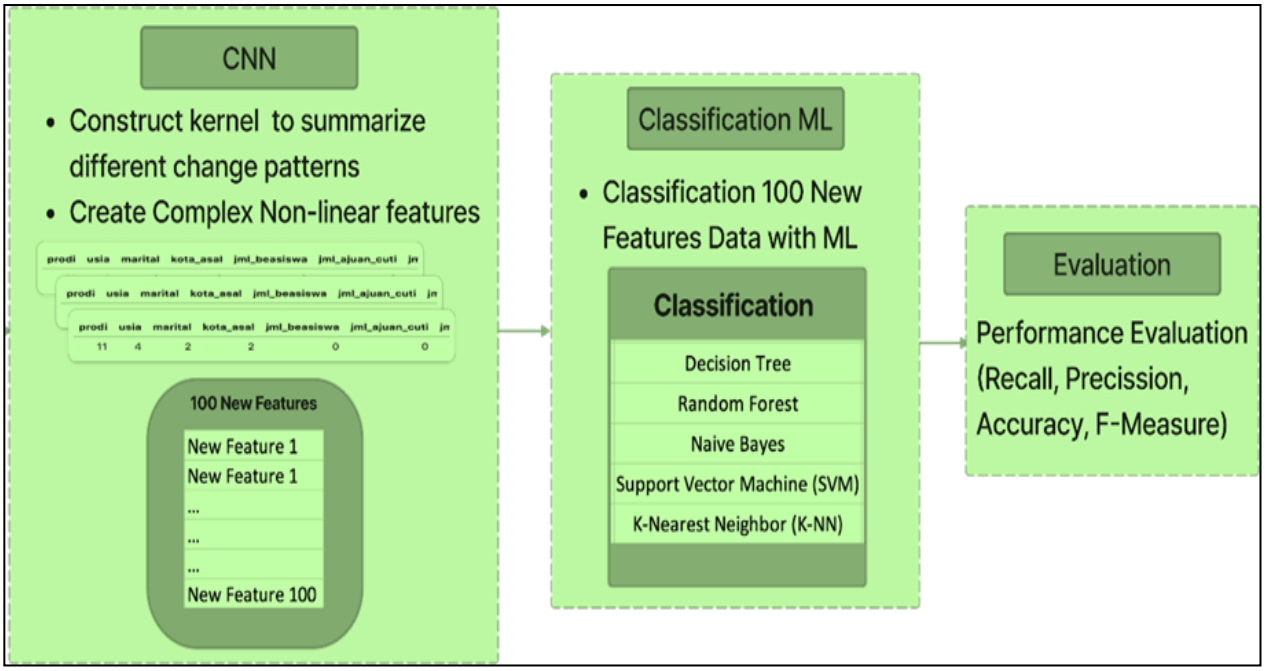

Fig. 5: Machine Learning Algorithm

Model: "model_1"

Layer (type)	Output Shape	Param #
conv2d_input (InputLayer)	[(None, 3, 4, 1)]	0
conv2d (Conv2D)	(None, 3, 4, 32)	1312
batch_normalization (BatchNo	(None, 3, 4, 32)	128
activation (Activation)	(None, 3, 4, 32)	0
conv2d_1 (Conv2D)	(None, 32, 4, 32)	3872
batch_normalization_1 (Batch	(None, 32, 4, 32)	128
activation_1 (Activation)	(None, 32, 4, 32)	0
conv2d_2 (Conv2D)	(None, 32, 4, 32)	40992
batch_normalization_2 (Batch	(None, 32, 4, 32)	128
activation_2 (Activation)	(None, 32, 4, 32)	0
conv2d_3 (Conv2D)	(None, 32, 4, 32)	40992
batch_normalization_3 (Batch	(None, 32, 4, 32)	128
activation_3 (Activation)	(None, 32, 4, 32)	0
max_pooling2d (MaxPooling2D)	(None, 32, 2, 16)	0
flatten (Flatten)	(None, 1024)	0
dense (Dense)	(None, 128)	131200
batch_normalization_4 (Batch	(None, 128)	512
activation_4 (Activation)	(None, 128)	0
dropout (Dropout)	(None, 128)	0
feature_dense (Dense)	(None, 100)	12900

Total params: 232,292
Trainable params: 231,780
Non-trainable params: 512

Fig. 6: Manual (File Eksperimen: 01-FeatureEng_Siadin_a2.Ipynb)

feauture_engg_data shape: (2293, 102)

	0	1	2	3	4	5	6	7	8	9	...	92	93	94	95	96	97	98	99	grade	nim
0	0.945863	0.131124	-0.589274	1.745187	-0.151498	0.444268	-0.432023	-1.617125	-0.813031	2.406016	...	-1.758027	1.709639	-0.212708	0.270016	-1.356441	0.470823	1.050641	-1.102458	0	23595
1	0.091682	-0.067778	-0.630636	0.008145	-1.008051	0.274363	-0.337376	0.416687	0.240006	0.893570	...	0.264423	0.604858	-0.846667	-0.401510	-0.371364	0.540907	0.326998	-0.361127	1	14249
2	0.192015	-0.464326	-0.614532	0.340080	-1.375439	0.204819	-0.012353	-0.284648	-0.384704	1.380224	...	-0.259371	1.133341	-0.675155	-0.004246	-0.782660	0.503713	0.433884	-0.369536	1	14350
3	-0.292109	-0.183514	-0.165619	-0.506382	-0.270095	-0.611406	0.039955	-0.712045	0.163402	1.695888	...	-0.248778	0.499986	-0.053213	0.307039	-0.310768	-0.606668	-0.269837	0.161650	1	16656
4	0.532819	0.391977	0.120263	-0.070715	0.136537	0.531833	0.037819	0.135637	-0.049361	0.692807	...	0.912891	-0.456833	-0.305826	-0.631459	-0.984662	0.076471	0.209569	0.320404	0	17076
5	0.925142	0.119603	1.851284	-1.813176	1.739573	0.414877	1.224937	0.078904	-0.368208	0.925756	...	1.388351	-0.781420	-0.448259	-0.345441	-2.446989	-0.219714	-0.701012	0.643477	1	17309
6	0.894703	0.675173	-0.051006	-0.536444	0.271438	0.613473	0.194302	0.298577	0.298965	0.336244	...	0.717968	-0.189463	-0.743676	-0.695498	-1.600466	0.048638	-0.192786	-0.295765	1	17459
7	0.566007	0.614056	1.897195	-1.589997	1.880853	-0.565920	1.427904	0.571449	-0.408568	1.090771	...	2.347176	-0.890110	0.029511	-1.259612	-1.502161	0.258081	-0.298442	0.647072	1	17561
8	0.747911	1.225748	-0.729559	-0.853543	0.797214	0.325384	0.180104	-1.872603	-0.324531	0.303313	...	-0.031446	0.161982	-1.537935	-0.681494	-0.439993	-0.660086	-0.199450	-0.392966	1	17701
9	0.220378	0.664507	0.082399	-0.414977	0.245052	-0.202754	0.397319	-0.727513	-0.445830	1.435740	...	-0.027529	0.517603	-0.016779	0.117569	-0.453499	-0.324775	0.064191	-0.153455	1	17888

10 rows × 102 columns

Fig. 7: Result of Extractione Fixture

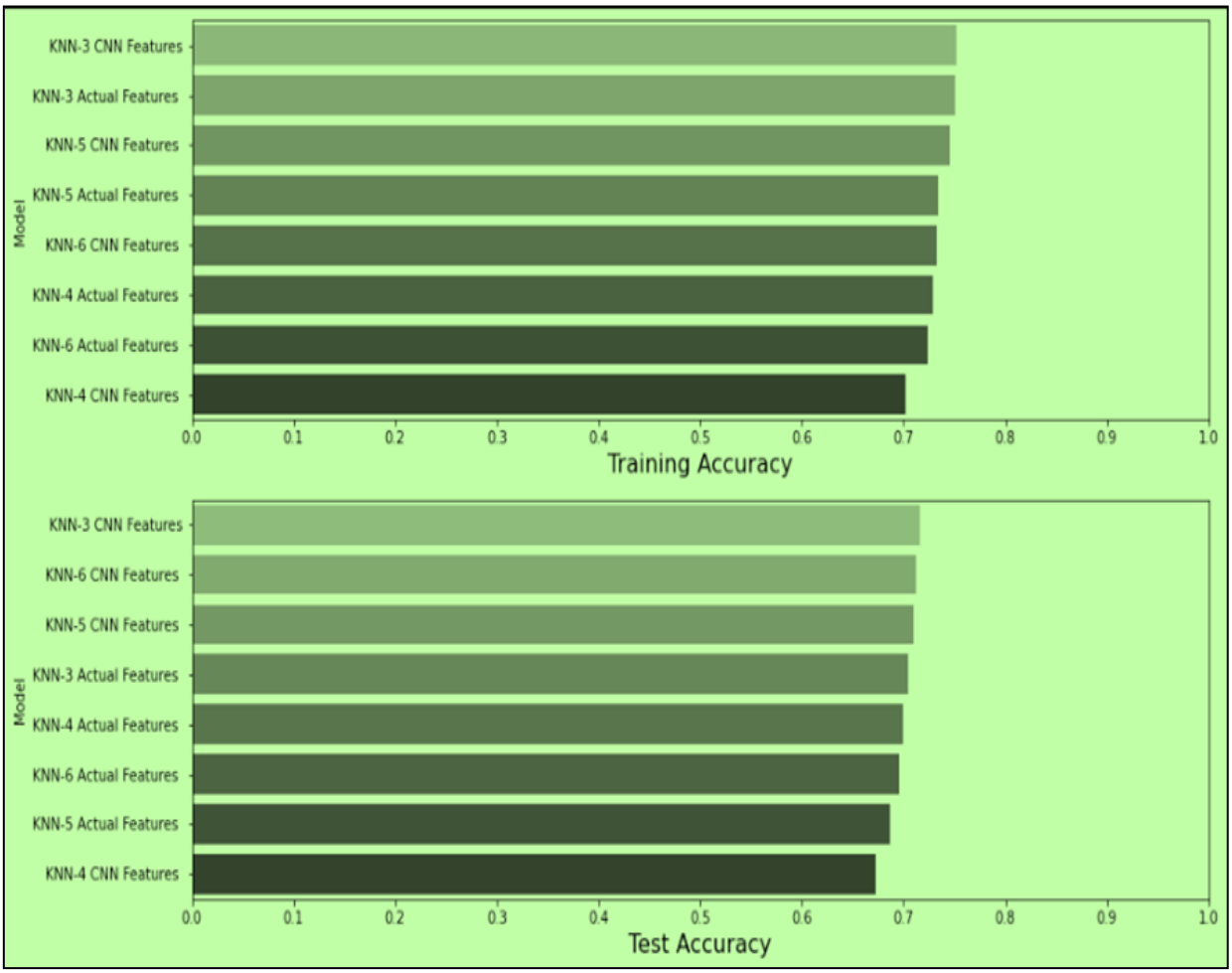

Fig. 8: File 100_Feature_Siadin2.csv

ISSN No:-2456-2165

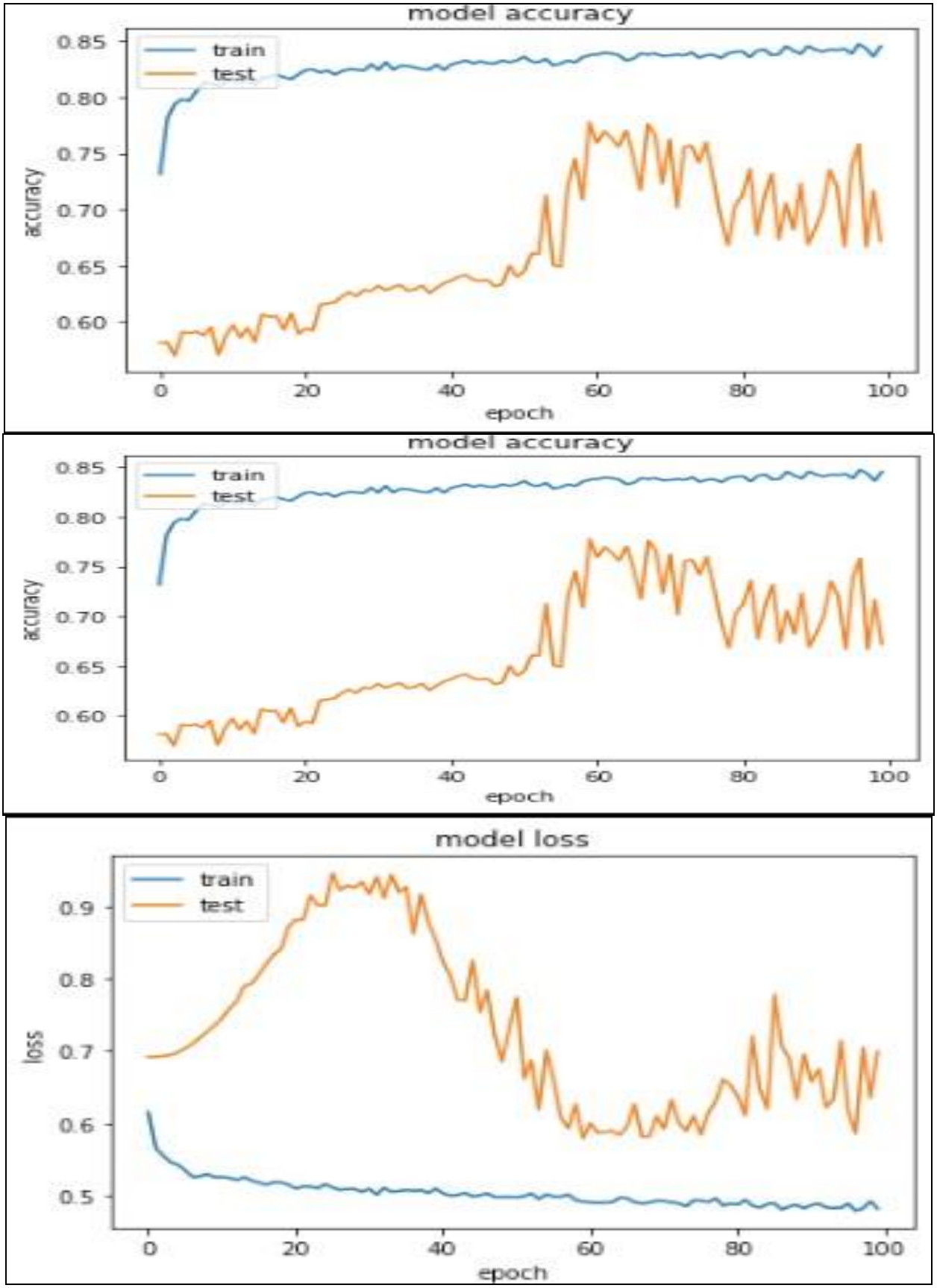

Fig. 9: Model Accuracy and Model Loss

feauture_engg_data shape: (2293, 130)

	0	1	2	3	4	5	6	7	8	9	...	120	121	122	123	124	125	126	127	grade	nim
0	-0.512210	1.088008	-0.116984	0.316776	-0.794252	-1.214166	-1.090983	0.139469	1.682442	1.473428	...	-1.463240	-0.893509	-0.251174	0.645127	0.641577	-0.305032	-1.025340	-0.281296	0	23595
1	0.142677	0.232225	-0.080689	0.422063	-0.192823	-0.121028	-0.167678	0.983777	1.007957	-0.387892	...	-0.464684	-0.207657	-0.272695	-0.528218	-0.297649	0.500515	0.431292	0.908169	1	14249
2	0.089541	0.574341	0.175454	0.458676	0.103949	-0.510401	-0.589814	0.842321	1.574392	0.378378	...	-1.114791	-0.568126	-0.092453	-0.040940	-0.455789	0.270388	-0.145014	0.346132	1	14350
3	-0.771723	-0.638912	0.032738	0.864526	-0.980690	-0.241999	0.142027	0.670698	-0.142912	-0.022366	...	-0.229398	-0.392066	1.131813	0.378191	1.123045	0.868740	0.477757	0.463972	1	16656
4	0.981043	0.231877	-0.992718	-0.207617	0.044727	0.288848	1.428707	-0.534004	-1.134766	0.022762	...	0.397682	0.730839	-0.144523	-0.480520	-0.409199	-0.220959	0.084485	-0.391283	0	17076
5	0.769950	0.228090	-0.371768	-1.860577	0.894118	0.770605	1.656164	-1.099850	-2.496315	1.432397	...	0.804921	1.276543	0.565855	0.133404	-1.549325	-0.870093	-0.576907	-1.726201	1	17309
6	0.584747	0.129143	-0.848724	-0.664081	0.233356	0.556354	0.905262	-0.159319	-0.941253	-0.190407	...	0.076800	1.371775	-0.127029	-0.578923	-0.202568	-0.380608	0.094863	0.077488	1	17459
7	0.661316	0.185858	-1.035413	-1.044456	0.455540	1.103668	1.719002	-0.608618	-1.962558	0.804563	...	0.731924	1.967411	0.091816	0.041247	-0.707052	-0.829744	0.122762	-0.819881	1	17561
8	-0.933302	-2.175135	-1.076449	0.841866	0.853569	-1.473366	0.174216	-1.419220	-2.790684	0.892105	...	0.653538	-0.352010	3.271122	0.993869	2.735859	0.635677	0.404208	-1.698491	1	17701
9	-0.961518	-0.907036	-0.790906	1.153867	-0.413131	-0.836842	-0.368063	0.395368	-0.599710	0.126136	...	0.219498	-0.626610	1.082720	0.983813	2.219875	0.971288	0.555904	-0.176437	1	17888

10 rows x 130 columns

Fig. 10: Figure Data

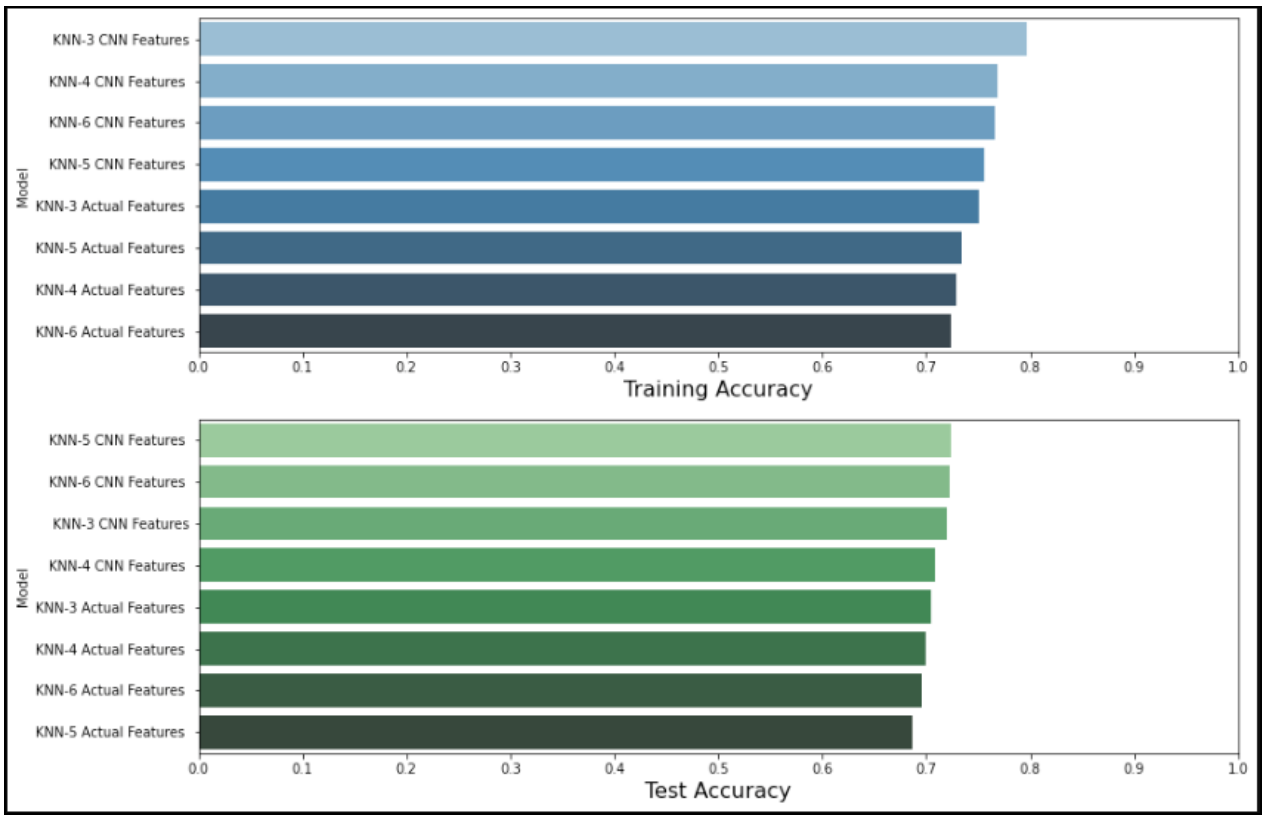

Fig. 11: End Results
(Data details in experimental file: 05-KNN-ML-Final.ipynb)

Shows the value of the results of Feature Extraction with CNN. Hyperparam results using the GA technique, overall accuracy has increased using K-NN for all n values: 3,4,5 and 6 (Proven Feature Extraction from tabular data represented in image form and processed using CNN algorithm uses the most appropriate parameters = Can be proven (SUCCESSFUL))

IV. CONCLUSION

Based on the research results, several conclusions were obtained as follows:

- The dataset used is a dataset of graduates from the A11 (Informatics Engineering – Bachelor's degree) study program at the Faculty of Computer Science, Dian Nuswantoro University, totaling 2293 records. The graduate data period is between the entry years 2012 to 2018

- Analysis of the Deep Learning Convolutional Network Method is carried out using CNN format data representation and CNN Architecture Hyperparameters

The experimental results show the value of the results of Feature Extraction with CNN. Hyparameter results using the Genetic Algorithm Technique, overall there is an increase in accuracy using K-NN for all values of n: 3,4,5 and 6 (Proven Feature Extraction from tabular data represented in the form of images and processed with the CNN algorithm using the most appropriate parameters = Can be proven (SUCCESSFUL)).

REFERENCES

[1]. Tekin, "Early Prediction of students' grade point averages at graduation: a data mining approach," Eurasian J. Educ. Res., vol. 14, no. 54, pp. 207–226, Feb. 2014.
[2]. M. K. Prof. Dr. Ir. Edi Noersasongko. (view May 2021). "Greetings from our Rector. [Online]. Available: https://dinus.ac.id/greetings.
[3]. ArunPriya, C. & Balasaravanan. 2012. An Efficient Leaf Recognition Algorithm for Plant Classification Using Support Vector Machine.
[4]. Goodfellow, I., Bengio, Y, & Courville, A. 2016. Deep Learning (Adaptive Computation and Machine Learning Series). The MIT Press.
[5]. Sugiyono, "Metode Penelitian Kuantitatif, Kualitatif dan R & D", Bandung : Alfabeta, 2012.
[6]. Salam, J. Zeniarja, and D. M. Anthareza, "Student Graduation Prediction Model using Deep Learning Convolutional Neural Network (CNN)," p. 5, 2022.
[7]. J. Zeniarja, A. Salam, and F. A. Ma'ruf, "Seleksi Fitur dan Perbandingan Algoritma Klasifikasi untuk Prediksi Kelulusan Mahasiswa," J. Rekayasa Elektr., vol. 18, no. 2, Jul. 2022, doi: 10.17529/jre.v18i2.24047.
[8]. N. M. Aszemi and P. D. D. Dominic, "Hyperparameter Optimization in Convolutional Neural Network using Genetic Algorithms," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 6, 2019, doi: 10.14569/IJACSA.2019.0100638.
[9]. F. Johnson, A. Valderrama, C. Valle, B. Crawford, R. Soto, and R. Nanculef, "Automating Configuration of Convolutional Neural Network Hyperparameters Using Genetic Algorithm," IEEE Access, vol. 8, pp. 156139– 156152, 2020, doi: 10.1109/ACCESS.2020.3019245.
[10]. Salam, J. Zeniarja, and D. M. Anthareza, "Student Graduation Prediction Model using Deep Learning Convolutional Neural Network (CNN)," p. 5, 2022.
[11]. J. Zeniarja, A. Salam, and F. A. Ma'ruf, "Seleksi Fitur dan Perbandingan Algoritma Klasifikasi untuk Prediksi Kelulusan Mahasiswa," J. Rekayasa Elektr., vol. 18, no. 2, Jul. 2022, doi: 10.17529/jre.v18i2.24047.