BAB 1 PENDAHULUAN

1.1 Latar Belakang

Stroke merupakan penyakit penyebab kematian yang menduduki peringkat kedua dan penyakit penyebab disabilitas ketiga di dunia. Selain itu, menurut data dari Riset Kesehatan Dasar (Riskesdas) Kementrian Kesehatan Republik Indonesia tahun 2018, meningkat jika dibandingkan data tahun 2013, yaitu dari 7% menjadi 10,9%. Jika dilihat dari keuangan, kasus stoke ini juga sangat berdampak. Menurut Badan Penyelenggara Jaminan Sosial (BPJS) Kesehatan, kasus stroke pada tahun 2016 sampai tahun 2018 sudah menghabiskan dana sekitar 4 triliun rupiah [1]. Oleh karena itu, diperlukan sebuah sistem yang dapat mendeteksi penyakit stroke lebih awal.

Sudah ada penelitian yang menggunakan algoritme *machine learning* dan *deep learning* untuk mendeteksi penyakit stroke. Pada penelitian sebelumnya, prediksi penyakit stroke dibuat dengan menggunakan algoritme seperti *Decision Tree*, *Gaussian Naïve Bayes*, *Random Forest*, *Expectation Maximization*, *Logistic Regression*, *K-Nearest Neighbors* (KNN), *Support Vector Machine* (SVM), dan *Deep Neural Network* (DNN) [2] [3] [4] [5] [6] [7] [8], namun model prediksi menggunakan *machine learning* belum dapat menghasilkan akurasi yang baik, sehingga diperlukan model untuk memprediksi penyakit stroke dengan menggunakan *deep learning* yang ditambahkan metode *regularization dropout* yang berfungsi untuk mencegah *overfitting*.

Pada Penelitian [2], dibandingkan beberapa metode *machine learning*, seperti *Logistic Regression*, *Decision Tree*, *Random Forest Classification*, *K-Nearest Neighbor* (KNN), *Support Vector Classification* (SVM), dan *Naïve Bayes*. *Dataset* pada penelitian ini diambil dari *website* Kaggle.com dengan nama *Stroke Prediction Dataset*, yang memiliki jumlah data sebanyak 5.110, di mana 249 stroke dan 4.861 tidak stroke. Penelitian ini juga memakai teknik *undersampling* untuk mengatasi *imbalanced class*. Hasil akurasi dan *recall* paling tinggi didapat oleh algoritme *Naïve Bayes*, yaitu dengan akurasi 82% dan *recall* 85,7%.

Penelitian [3] dibandingkan beberapa metode *machine learning* seperti *Decision Tree*, *Logistic Regression*, dan *Random Forest*. Data didapatkan dari *dataset* di situs Kaggle.com dengan nama *Healthcare stroke Patients in Python*, yang terdiri dari 12 kolom dan 62.001 baris. *Random Forest* meraih akurasi dan

recall tertinggi, yaitu 99,98% dan 99%. Hasil dari penelitian ini terlihat mengalami *overfitting* yang sangat tinggi.

Penelitian lain [4] membandingkan algoritme lain, yaitu *Decision Tree*, *Expectation Maximization*, *Random Forest*, *Gaussian Naive Bayes*, dan *Deep Neural Network* (DNN). Penelitian ini juga memakai *Principal Component Analysis* (PCA) sebagai teknik *feature extraction*. Data didapat dari banyak rumah sakit di Banglore dan *medical center*, dengan total sebesar 1.500 data. Dalam penelitian ini disebutkan bahwa model yang dibuat dengan DNN dan *feature extraction* PCA mendapatkan hasil akurasi dan *recall* terbaik, yaitu 86,42% dan 74,89%, namun dalam penelitian ini juga disebutkan bahwa DNN memiliki kelemahan, yaitu waktu *training* yang lambat sehingga kita harus meningkatkan performa.

Penelitian [5] bertujuan mencari algoritme yang paling cocok untuk kasus *dataset* yang sangat besar, sekitar 800 ribu data. Penelitian ini membandingkan DNN, *Gradient Boosting Decision Tree* (GBDT), *Logistic Regression*, dan SVM. Data penelitian didapat dari National Health Insurance Research Database (NHIRD) dan hanya memakai 2.007 fitur dari total keseluruhan 7.932 fitur. Model DNN mendapat hasil terbaik dengan akurasi 87,3%, *recall* 84,5%, dan AUC 91,5%

Penelitian [6] bertujuan membandingkan AUC antara ANN tanpa *scaling* dengan ANN menggunakan bermacam-macam *scaling* (*normalizer*, *min-max*, *standard*, dan *robust*), SVM, XGB, *Binary Logistic Regression*. Hasil terbaik didapat oleh model ANN tanpa *scaling*, dengan akurasi 87,8%, *recall* 96,7%, ROC 84%.

Penelitian [7] bertujuan mencari kombinasi hyperparameter terbaik untuk mendapatkan akurasi tertinggi pada model DNN. Data didapatkan dari Imam Khomeini Hospital, Ardabil, Iran, dengan jumlah 332 pasien. Penelitian ini melakukan 81 percobaan terkait kombinasi *activation function*, *hidden layer*, *epoch*, *momentum*, dan *learning rate*. Hasil terbaik diraih dari kombinasi *activation function* tanh, *hidden layer* berjumlah 10, *epoch* berjumlah 400, *momentum* sebesar 0,5, *learning rate* 0,1, dengan akurasi sebesar 99,5%, *recall* 98%, dan ROC area 97%.

Penelitian lain [8], dilakukan pembuatan model menggunakan DNN dan PCA agar dapat mencari variabel yang berperan paling penting dalam kasus penyakit stroke. Data didapat dari *Korean National Hospital Discharge In-depth Injury Survey* (KNHDS). KNHDS mengambil data dari *Korea Centers for Disease*

Control and Prevention (KCDC), yang dikumpulkan dari tahun 2013 hingga tahun 2016. Penelitian ini menggunakan 15.099 data dan 11 variabel. Model DNN yang didapat dari penelitian ini mendapat akurasi 84,03%, *recall* 64,32%, dan AUC 83,48%.

Penelitian [9] bertujuan untuk membandingkan 10 *dataset* yang semuanya bersifat *imbalanced class*, dan hasilnya 6 data mendapatkan G-Mean yang terbaik dengan teknik *cost-sensitive learning with moving threshold* dengan metode pengukuran ROC *curve*, jika dibandingkan dengan metode *random oversampling*, *random undersampling*, SMOTE, *cost-sensitive learning with moving threshold* dengan metode pengukuran *imbalance ratio*.

Pada penelitian ini, akan digunakan metode *Deep Neural Network* (DNN) dengan memperhatikan *dropout* [10] [11], *cost-sensitive learning*, dan *probability tuning*. Teknik *dropout* digunakan untuk mengatasi kelemahan DNN, yaitu mudah mengalami *overfitting* [7] dan waktu *training* yang lambat [4] [8]. *Cost-sensitive learning* dan *probability tuning* digunakan karena *dataset* yang digunakan bersifat *imbalanced*.

1.2 Rumusan Masalah

Berikut adalah rumusan masalah yang akan dibahas di dalam penelitian ini.

1. Berapa kombinasi terbaik dari *hyperparameter* seperti *learning rate*, *hidden layer*, *epoch* dan *activation function* untuk memprediksi seseorang terkena stroke?
2. Bagaimana pengaruh *dropout* dalam mengatasi *overfitting* pada metode DNN untuk memprediksi seseorang terkena stroke?
3. Bagaimana pengaruh *cost-sensitive* dan *probability tuning* dalam mengatasi *dataset* yang bersifat *imbalanced* pada metode DNN untuk memprediksi seseorang terkena stroke?
4. Berapa nilai ROC terbaik pada model DNN untuk memprediksi seseorang terkena stroke?

1.3 Tujuan Penelitian

Berikut adalah tujuan penelitan dalam penelitian ini.

1. Mengetahui pengaruh *dropout* dalam mengatasi *overfitting* dengan metode DNN.
2. Mengetahui pengaruh *cost-sensitive* dan *probability tuning* dalam mengatasi *dataset* yang bersifat *imbalanced* pada metode DNN.
3. Mengetahui nilai ROC terbaik pada model DNN.

1.4 Batasan Masalah

Agar penelitian ini menjadi lebih terarah, maka masalah yang akan dibahas akan dibatasi sebagai berikut.

1. Dataset yang digunakan berasal dari Kaggle, dengan judul *Cerebral Stroke Prediction-Imbalanced Dataset*.
2. *Overfitting* atau tidaknya suatu model akan dilihat dari *learning curve*.
3. Model akan dilihat performanya dari nilai *Receiver Operating Characteristic* (ROC) *Curve*.

1.5 Kontribusi Penelitian

Kontribusi yang diberikan pada penelitian ini adalah sebagai berikut.

1. Melakukan pengujian apakah metode DNN cocok untuk prediksi penyakit stroke pada seseorang.
2. Melihat seberapa berpengaruh *overfitting* pada *dataset* tabular dengan algoritme DNN terhadap akurasi data *testing*.
3. Melakukan pengujian apakah *dropout* dapat benar-benar mengatasi *overfitting*.

1.6 Metodologi Penelitian

Penelitian ini dibuat dengan metode penelitian sebagai berikut.

1. Studi Literatur

   Penulisan tugas akhir ini dimulai dengan melakukan studi kepustakaan yaitu dengan cara mengumpulkan bahan-bahan referensi seperti jurnal penelitian, *paper*, dan buku terkait dengan topik.
2. Eksplorasi Dataset

   Pada tahap ini penulis akan mempelajari isi dan karakteristik dari dataset *Cerebral Stroke Prediction-Imbalanced Dataset* yang akan digunakan untuk memprediksi kemungkinan penyakit stroke pada seseorang.
3. Analisis Masalah

   Pada tahap ini akan dilakukan analisis permasalahan yang ada berdasarkan batasan masalah yang sudah dibuat.
4. Perancangan dan Implementasi Algoritme

   Pada tahap ini akan dilakukan pembuatan model dengan algoritme DNN dan *dropout*.
5. Pengujian

   Pada tahap ini akan dilakukan pengujian terhadap hasil akurasi prediksi stroke

dengan cara hasil akurasi dan ROC *curve* akan dibandingkan antara algoritme DNN dengan teknik *regularization dropout* dan DNN yang tidak menggunakan *dropout*.

6. Dokumentasi

   Pada tahap ini akan dilakukan dokumentasi hasil analisis dan implementasi secara tertulis dalam bentuk laporan tugas akhir.

1.7 Sistematika Pembahasan

Penelitian ini dibuat dengan sistematika sebagai berikut.

BAB 1 PENDAHULUAN: Bab ini berisi latar belakang, rumusan masalah, tujuan penelitian, batasan masalah, kontribusi penelitian, metodologi penelitian, dan sistematika pembahasan

BAB 2 LANDASAN TEORI: Bab ini berisi penjelasan dasar mengenai teori yang mendukung untuk implementasi penelitian ini.

BAB 3 METODOLOGI PENELITIAN: Bab ini berisi analisis algoritme DNN dengan *regularization dropout* dan *handling imbalanced class* menggunakan *cost-sensitive learning* dan *probability tuning* untuk membangun model prediksi orang terkena penyakit stroke.

BAB 4 IMPLEMENTASI DAN PENGUJIAN: Bab ini berisi implementasi dan pengujian dari algoritme DNN, *regularization dropout*, *cost-sensitive learning*, dan *probability tuning* terhadap *dataset* stroke, melihat performa model dengan ROC *curve*, dan melihat *overfitting* atau tidaknya suatu model menggunakan *learning curve*.

BAB 5 KESIMPULAN DAN SARAN: Bab ini berisi kesimpulan dari penlitian yang dilakukan berdasarkan hasil dari pengujian dan saran untuk penelitian di waktu mendatang.