1. PENDAHULUAN

1.1. Latar Belakang

Gastroesophageal Reflux Disease (GERD) adalah gangguan berupa regurgitasi isi lambung yang menyebabkan *heartburn* dan gejala lain[1]. Angka Prevalensi dari GERD berbeda di tiap daerah. Angka tersebut diantaranya amerika selatan yang bernilai 23%, Amerika Utara 18.1%-27.8%, Australia 11.6%, Eropa 8.8%-25.9%, dan Asia Timur 2.5% - 7.8%. Pada tahun 1997, terjadi peningkatan prevalensi GERD dari 5.7% menjadi 25.18% di indonesia [2]. Gejala yang ditimbulkan oleh GERD cenderung lebih umum. Oleh karena itu, gejala yang timbul belum tentu bisa digolongkan sebagai GERD terlebih jika tidak terdapat gejala khas GERD yang diketahui [3].

Deteksi dini GERD diperlukan karena GERD diketahui sebagai penyebab radang tenggorokan posterior dan batuk kronis [4]. Selain itu, GERD juga dianggap sebagai faktor risiko *Idiopathic pulmonary fibrosis (IPF)* [5]. Saat ini, deteksi GERD pada umumnya dilakukan dengan metode *Proton Pump Inhibitor* (PPI) karena metode tersebut sederhana dan tidak memerlukan banyak biaya [6]. Akan tetapi, dalam jangka panjang metode ini memiliki efek samping bagi penggunanya seperti alergi, demensia, hingga kanker lambung [7]. Oleh karena itu dibutuhkan metode deteksi dini GERD lain yang relatif aman dan efisien.

Deteksi penyakit dengan *machine learning* merupakan salah satu peminatan yang luar biasa di komunitas biomedis. Hal tersebut dikarenakan *machine learning* dapat meningkatkan sensitifitas dan/atau spesifitas deteksi dan diagnosis penyakit [8]. Deteksi penyakit GERD dengan *machine learning* sudah pernah dilakukan oleh Huang dkk pada tahun 2015 dengan kernel *Hierarchical Heterogeneous Descriptor Fusion* Support Vector Machine (HHDF-SVM) pada gambar endoskopi dari *esophageal-gastric junctions* pasien. Penelitian tersebut menghasilkan akurasi senilai 93.2% [9].

Selain pada gambar, deteksi penyakit dengan *machine learning* juga bisa dilakukan pada sumber textual seperti pada tahun 2020, Amin melakukan penelitian dengan *term frequency-inverse document frequency* (*TFIDF*) dalam deteksi penyakit dengue melalui tweets menghasilkan akurasi diantara 80.81%-92.88% dengan metode klasifikasi yang berbeda beda dimana metode *Support Vector*

Machine (SVM) memiliki akurasi yang paling kecil dan *Long-Short Term Memory* (LSTM) memiliki akurasi yang tertinggi[10]. Pada tahun 2018 juga Bhasuran dkk melakukan penelitian dengan metode *joint ensemble learning* dalam deteksi penyakit gen dengan menggunakan *biomedical literature* telah dilakukan. Penelitian tersebut diterapkan pada beberapa *corpus* yang berbeda dan dibandingkan dengan beberapa metode *text mining* yang berbeda, menghasilkan nilai *f1-score* pada selang 83.93-87.39% dimana nilai tersebut masih berada diatas nilai metode *text mining* lainnya kecuali pada *corpus polysearch* [11].

Pada tahun 2020, penelitian tentang deteksi COVID-19 telah dilakukan oleh Khanday dkk dengan berbagai metode *machine learning* pada data teks klinis. Data teks klinis divektorisasi dengan TFIDF dan klasifikasi oleh beberapa metode *machine learning*. Penelitian ini menghasilkan nilai akurasi diatas 90% untuk semua metode *machine learning* yang diterapkan [12].

Pada tugas akhir ini, penulis akan membangun model deteksi penyakit GERD dengan menggunakan *feature importance* untuk seleksi fitur dan metode *ensemble* untuk membangun model prediksi berdasarkan ulasan obat yang dikonsumsi dari pasien. *Feature importance* digunakan karena terbukti memberikan cara yang lebih baik dalam mengukur relevansi fitur [13], sedangkan metode *ensemble learning* digunakan karena terbukti meningkatkan performansi dari model berbasis teks[14]. Selain itu, metode ensemble juga dapat menghasilkan model yang lebih baik pada dataset yang banyak [15]. Algoritma yang akan digunakan untuk membangun model prediksi pada ensemble learning adalah algoritma *random forest* dan *adaptive boosting (AdaBoost)*

1.2. Perumusan Masalah

Rumusan masalah yang akan dibahas dalam tugas akhir ini adalah :

1. Bagaimana proses seleksi data dengan *feature importance* ?
2. Bagaimana membangun model dengan menggunakan metode *ensemble* untuk deteksi GERD ?
3. Bagaimana performa metode *ensemble* dalam melakukan deteksi GERD berbasis teks *?*

1.3. Tujuan

Pada penulisan Tugas Akhir ini, tujuan yang ingin dicapai sebagai berikut :

1. Melakukan proses seleksi data dengan *feature importance*
2. Membangun model dengan menggunakan metode *ensemble* untuk deteksi GERD.
3. Mengetahui performa metode *ensemble* dalam melakukan deteksi GERD berbasis teks.

1.4. Batasan Masalah

Dataset yang digunakan dalam Tugas Akhir ini adalah dataset dengan judul "UCI ML Drug Review dataset" dan dapat diakses pada link https://www.kaggle.com/jessicali9530/kuc-hackathon-winter-2018

1.5. Rencana Kegiatan

Berikut adalah rencana kegiatan yang akan dilakukan terkait penyelesaian masalah pada tugas akhir.

1. Studi Literatur

   Pada tahap ini, aspek-aspek penting pada penelitian ini dipelajari lebih dalam melalui literatur yang ada. Literatur tersebut dapat berbentuk buku, jurnal, atau sumber informasi lainnya yang dapat dipertanggungjawabkan isinya. Tahap ini diperlukan guna mengetahui bagaimana penelitian ini dapat diselesaikan.

2. Preparasi Data

   Pada tahap ini, dataset yang digunakan pada penelitian dipersiapkan agar proses selanjutnya dapat berjalan lebih baik. Dataset akan masuk kedalam suatu proses bernama *Pre-Processing*. Proses ini akan membersihkan dataset dari berbagai kemungkinan kesalahan pada model yang dibangun contohnya seperti data kosong pada dataset. Selanjutnya, dataset akan diekstrak dengan *Text-Processing*.

3. Analisis dan Perancangan Model

   Rancangan Model Prediksi akan dihasilkan pada tahapan ini. Rancangan tersebut dibangun berdasarkan analisis terhadap aspek-aspek yang diketahui pada tahapan studi literatur.

4. Pembangunan Model Prediksi

   Pada tahap ini, model prediksi akan dibangun menggunakan metode *ensemble*.

5. Evaluasi Model dan Analisis

   Pada tahap ini, model akan dievaluasi dan dianalisis berdasarkan berbagai parameter validasi model seperti *confusion matrix*, *accuracy*, *precision*, *recall*, *specificity* dan *f-1 score*.

6. Penulisan Laporan

   Hasil dari tahapan-tahapan sebelumnya akan dilaporkan sebagai dokumentasi. Dokumentasi berisi penjelasan mengenai keseluruhan proses, hasil, serta analisis yang telah dilakukan.

1.6. Jadwal Kegiatan

Tugas akhir ini dibuat sesuai dengan timeline pada Tabel 1 sebagai berikut:

Tabel 1. Timeline

Kegiatan	**Bulan**					
	1	2	3	4	5	6
Studi Literatur						
Preparasi Data						
Analisis dan Perancangan Model						
Pembangunan Model Prediksi						
Validasi Model dan Analisis						
Penulisan Laporan						

*Keterangan: shading warna *grayscale*