BAB I PENDAHULUAN

1.1 Latar Belakang

Survey yang dilakukan lima tahun sekali pada tahun 2018 dan 2023, menurut Survey Kesehatan Indonesia (SKI) dan laporasn riset kesehatan dasar (riskesdas) yang dilakukan oleh Kementerian Kesehatan Republik Indonesia menyebutkan bahwa ada enam jenis penyakit tidak menular yang apabila tidak ditangani dengan serius akan menjadi penyakit kronis [1], [2]. Penyakit-penyakit tersebut yaitu diabetes mellitus (DM), kanker, jantung (CVD), gagal ginjal kronis (GGK), stroke, dan hipertensi (HT). Penyakit kronis tersebut membutuhkan biaya perawatan yang besar dan banyak diderita oleh masyarakat di Indonesia [2]. Berdasarkan SKI, penyakit kronis tidak menular tersebut mengalami peningkatan penderita seiring dengan bertambahnya usia [2].

Untuk mencegah peningkatan jumlah penderita di tahun-tahun mendatang, diperlukan upaya *preventif* yang terintegrasi antara teknologi dan tenaga medis. Salah satu langkah yang dapat dilakukan adalah mendorong masyarakat untuk menjalani pemeriksaan kesehatan secara rutin. Selanjutnya, hasil pemeriksaan *general check-up* (GCU) dapat dianalisis untuk mengidentifikasi faktor risiko, dengan keterlibatan aktif dari tenaga medis. Selain itu, pemanfaatan model *machine learning* dapat digunakan sebagai alat bantu dalam proses prediksi diagnosis, guna mengidentifikasi potensi penyakit tunggal maupun komorbid.

Seiring dengan perkembangan teknologi saat ini, tentu model *machine learning* sudah banyak digunakan untuk memprediksi berbagai jenis penyakit berdasarkan data riwayat hasil pemeriksaan kesehatan seseorang (GCU) [3]–[12]. Hasil yang diperoleh dari beberapa penelitian tersebut menyebutkan bahwa ada informasi dari hasil pemeriksaan yang dapat meningkatkan resiko seseorang terkena penyakit [3]–[5]. Meski teknologi *machine learning* sudah banyak digunakan dalam prediksi penyakit, peran dokter tetap menjadi kunci dalam proses diagnosis klinis.

Saat ini, proses diagnosis penyakit oleh dokter mengandalkan wawancara medis dengan pasien, pemeriksaan fisik, serta pemeriksaan penunjang, yang kemudian diproses melalui *clinical reasoning* berbasis pengalaman dan keahlian untuk menyusun diagnosis yang paling mungkin terjadi pada seorang pasien. Namun, di tengah kompleksitas dan

volume data medis yang semakin besar dan heterogen, dokter memerlukan dukungan sistem berbasis *machine learning* untuk meningkatkan akurasi, efisiensi, dan keandalan dalam pengambilan keputusan klinis. Sehingga, tantangan utamanya adalah mengintegrasikan antara pengetahuan klinis seorang dokter dan *machine learning* secara sinergis agar hasil analisis tidak hanya akurat secara statistik, tetapi juga kredibel bagi para ahli.

Selanjutnya, beberapa penelitian tentang prediksi penyakit kronis yaitu Fitriyani, Norma Latif, dkk melakukan prediksi penyakit hipertensi dan diabetes [13]. **Namun**, pada penelitian tersebut data yang digunakan berasal dari empat jenis data untuk ditentukan variabel mana yang beririsan dalam penyakit diabetes dan hipertensi, dalam penentuan variabel yang beririsan penelitian tersebut tidak melibatkan pengetahuan dokter. Beberapa penelitian melakukan prediksi penyakit menggunakan data yang juga mencakup variabel penyakit lainnya seperti diabetes, serangan jantung, hipertensi, dan stroke [14]–[22]. Penelitian lain membahas tentang kemungkinan seseorang terkena satu penyakit [23], [24]. **Namun,** pada penelitian-penelitian tersebut hanya berfokus pada 1 jenis penyakit dan tidak menjelaskan *output* seberapa besar kemungkinan terkena penyakit yang diderita.

Christel, dkk melakukan tinjauan komprehensif terhadap upaya-upaya dalam mengintegrasikan pengetahuan medis ke dalam *machine learning* [25]. Prosesnya mencakup *preprocessing* data, rekayasa fitur, pelatihan model, dan evaluasi output. Studi tersebut, selanjutnya mengeksplorasi signifikansi dan dampak dari integrasi melalui studi kasus tentang prediksi diabetes. Studi tersebut, pengetahuan klinis yang mencakup aturan, jaringan kausal, interval, dan rumus, diintegrasikan pada setiap tahap *machine learning*. **Namun**, studi tersebut masih memiliki sejumlah keterbatasan yang belum sepenuhnya dilakukan, antara lain keterbatasan generalisasi karena hanya diuji pada satu dataset publik (PIMA Diabetes), belum ditanganinya permasalahan data medis seperti imbalanced class, dan bentuk integrasi domain knowledge dengan cara mengubah loss function dan voting scheme menggunakan aturan.

Lebih lanjut, beberapa penelitian terkait penanganan *preprocessing* pada data. Su, Yifei, dkk memiliki isu utama penelitian yaitu model prediksi dalam kemampuan generalisasi masih rendah yang diakibatkan *imbalanced* dataset. Proses penanganan *imbalanced* menggunakan model SMOTE. Penelitian tersebut melakukan adaptasi

dengan mengelompokkan data berdasarkan kategori usia menggunakan teknik *feature compensasition* [26]. **Namun**, pada penelitian tersebut untuk teknik adaptasi tidak menerapkan pengetahuan dokter dalam proses pembuatan data sintetisnya. J. A. Castellanos-Garzón menyelesaikan permasalahan terkait *maximum rule* dan *intersection rule* pada dataset penyakit diabetes mellitus, kanker, dan jantung [27]. *Rule* yang dihasilkan didapat dari model klasifikasinya. **Namun**, kedalaman *rule* yang dihasilkan dari algoritma tersebut cenderung tetap (tidak bisa berkurang atau bertambah).

Sergio, dkk menggunakan data dari rumah sakit di Jepang, dengan dua skenario penanganan data: mengisi nilai yang hilang dan mengabaikannya [4]. Zou, Quan, dkk memproses data rumah sakit di China dengan menghapus nilai *null* dan mereduksi dimensi [28]. Azrar, Amina, dkk menggunakan data sekunder dari repositori terbuka, menangani nilai hilang dengan *mean*, serta mengonversi data numerik menjadi kategorik [29]. Wu, Han, dkk menggunakan data publik dengan melakukan penanganan pada data yaitu konvert numerik ke nominal, mengganti *missing value* menggunakan *mean*, dan normalisasi data menggunakan *z-score* [30]. Maniruzzaman, Md, dkk mengisi nilai hilang dan mengganti outlier menggunakan median dan membandingkan beberapa metode seleksi fitur [31]. **Namun**, secara keseluruhan untuk penanganan *imbalanced* data dan model prediksi masih belum ada yang menerapkan *domain knowledge* ke dalam proses tersebut.

Bedasarkan pemaparan sebelumnya, maka terdapat dua permasalahan dalam prediksi penyakit kronis. Permasalahan pertama yaitu dokter membutuhkan *insight* tambahan untuk melakukan terapi preventif kepada seseorang yang berpotensi mengalami penyakit kronis di masa depan. Permasalahan kedua muncul ketika data medis yang digunakan tidak dipersiapkan dengan baik yang ditandai dengan adanya nilai *null*, karakter nilai pada data yang tidak diperlukan, *outlier*, dan ketidakseimbangan jumlah pasien positif dan negatif (*imbalanced*). Beberapa permasalahan tersebut saat ini banyak diteliti dengan menggunakan pendekatan statistik dan *machine learning* [13], [32], [33]. Oleh karena itu, masih terdapat peluang untuk menggunakan pendekatan baru dalam menangani permasalahan data serta memodifikasi model prediksi *machine learning* berdasarkan *domain knowledge*.

1.2 Perumusan Masalah

Berdasarkan latar belakang di atas, berikut merupakan *research question* (RQ) penelitian ini:

1. Penelitian sebelumnya dalam menangani data yang tidak seimbang umumnya menggunakan pendekatan seperti pembuatan data sintetis.

   RQ-1: Pendekatan seperti apa yang digunakan untuk menangani data GCU yang imbalanced?
2. Beberapa penelitian yang telah ada untuk prediksi penyakit kronis biasanya berfokus pada prediksi 1 jenis penyakit. Berdasarkan penelitian terdahulu penanganan data yang *imbalanced* dan model prediksi menggunakan statistik atau *machine learning*.

   RQ-2: Bagaimana menghasilkan model yang memuat *domain knowledge* untuk memprediksi seseorang terkena satu atau beberapa penyakit?

1.3 Tujuan

Tujuan penelitian ini dapat ditentukan dengan merujuk pada masing-masing rumusan permasalahan. Berikut merupakan tujuan penelitian ini:

1. Meningkatkan kualitas data GCU melalui pembuatan data sintetis (**RQ-1**) .
2. Menerapkan *domain knowledge* pada penanganan *preprocessing* dan model prediksi yang dapat berpotensi meningkatkan hasil performansi prediksi (**RQ-2**).

1.4 Manfaat Penelitian

Manfaat penelitian ini dapat ditentukan dengan merujuk pada masing-masing tujuan penelitian. Berikut merupakan manfaat penelitian ini:

1. Hasil prediksi menjadi lebih akurat sehingga, dapat meningkatkan keandalan model dalam mengidentifikasi individu dengan risiko tinggi secara lebih dini, serta memperkuat dasar pengambilan keputusan dalam tindakan preventif medis (**RQ-1**).
2. Prediksi tidak hanya memanfaatkan *machine learning* saja tetapi juga melibatkan informasi klinis dari seorang pakar medis (**RQ-2**).

1.5 Dampak Penelitian

Penelitian ini tentunya memiliki beberapa dampak dari sisi pasien, pakar kesehatan, dan pihak asuransi. Berikut merupakan beberapa manfaat tersebut:

1.5.1 Sisi Pasien

Penelitian ini diharapkan memberikan berbagai manfaat bagi pasien. Berikut adalah beberapa manfaat yang dapat diperoleh dari perspektif pasien:

1. Supaya pihak pasien dapat melakukan monitoring kesehatan dirinya sendiri setiap saat dengan harapan tingkat penderita penyakit kronis menurun.
2. Menekan biaya pengeluaran untuk membeli obat-obatan.
3. Memberikan pasien wawasan tentang faktor gaya hidup yang memengaruhi kesehatan mereka, misalnya riwayat merokok atau kurangnya aktivitas fisik.

1.5.2 Sisi Pakar Kesehatan

Penelitian ini diharapkan memberikan berbagai manfaat bagi pakar kesehatan. Berikut adalah beberapa manfaat yang dapat diperoleh dari perspektif pakar kesehatan:

1. Dokter dapat membuat perencanaan perawatan yang lebih tepat dan mengarahkan pasien ke intervensi yang paling relevan sesuai dengan profil risiko individu.
2. Membantu dokter dalam pengambilan keputusan klinis yang lebih terinformasi.

1.5.3 Sisi Pihak Asuransi

Penelitian ini diharapkan memberikan berbagai manfaat bagi pihak asuransi. Berikut adalah beberapa manfaat yang dapat diperoleh dari perspektif pihak asuransi:

1. Membantu perusahaan asuransi dalam menilai risiko kesehatan calon pelanggan secara lebih akurat, sehingga memungkinkan untuk penentuan premi yang lebih tepat.
2. Asuransi dapat mengoptimalkan biaya klaim di masa depan karena penanganan penyakit yang lebih terkontrol.
3. Asuransi dapat menawarkan paket atau program kesehatan yang sesuai dengan risiko kesehatan individu, misalnya asuransi khusus bagi mereka yang berisiko tinggi terkena diabetes atau hipertensi.

1.6 Hipotesis

Berikut merupakan hipotesis beserta premis yang dihasilkan penelitian ini:

1. Hipotesis 1

Premis 1:

Dritsas, Elias, dkk melakukan prediksi stroke dengan menggunakan *stacking ensemble* (*naive bayes*, *random forest*, *RepTree*, dan J48) [15]. Wang, Qian, dkk dan N.G. Ramadhan, dkk melakukan prediksi diabetes mellitus dengan menggunakan model berbasis *ensemble learning* yaitu *random forest* [32], [34]. Cai, Tongan, dkk melakukan prediksi kanker dengan menggunakan model *ensemble learning* yaitu *random forest* dan *xtreme gradient boosting* yang dibandingkan dengan model *bayes* [35].

Premis 2:

Wang, Qian, dkk melakukan prediksi diabetes mellitus dengan menangani *missing value* menggunakan naïve bayes [34]. Qi, Huitao, dkk melakukan prediksi kanker pada tahap awal dengan menerapkan nilai *mean* untuk menangani *missing value* [14]. NG. Ramadhan, dkk menjelaskan bahwa hasil prediksi penyakit diabetes mellitus dapat meningkatkan *f1-score* sebesar 25% dengan menangani *imbalanced* data [32]. Metode *imbalanced* yang digunakan yaitu *random oversampling* [32]. Xiao, Yawen dalam hasil penelitiannya menganalisis perbedaan hasil prediksi penyakit kanker dengan permasalahan *imbalanced* data dapat meningkatkan *f1-score* sebesar 18%. Metode *imbalanced* yang digunakan adalah *Wasserstein Generative Adversarial Networks* (*oversampling*) [36].

Pengaruh *imbalanced* data juga dijelaskan pada penelitian Fitriyani, Norma Latif, dkk yang mana hasil prediksi penyakit diabetes mellitus dan hipertensi menunjukkan dapat ditingkatkan lebih dari 15% jika dilakukan penanganan *imbalanced* data. Metode *imbalanced* yang digunakan adalah SMOTETomek (*oversampling*) [13]. López-Martínez, Fernando, dkk pada penelitiannya menyebutkan bahwa dampak *imbalanced* data terhadap hasil prediksi penyakit hipertensi dapat meningkatkan sebesar 30%. Metode *imbalanced* yang digunakan yaitu SMOTE (*oversampling*) [33].

Hipotesis:

Ensemble learning memiliki kemampuan dalam memprediksi kemungkinan seseorang terkena salah satu atau beberapa penyakit kronis dengan tingkat F1-Score melebihi rata-rata nilai F1-Score mencapai 71%, dan (2) penanganan permasalahan pada data *GCU* yang *imbalanced* dapat mempengaruhi hasil F1-score minimal 15%.

2. Hipotesis 2

Premis 1:

Penelitian lain dalam hal melakukan prediksi dan menangani data yang kotor biasanya menggunakan metode statistik atau *machine learning* [14], [26], [34].
Premis 2:

Pada penelitian lain dalam menghasilkan kedalaman *rule* untuk prediksi cenderung tetap [27].

Hipotesis:

Pengetahuan dokter diintegrasikan dalam proses *preprocessing*, khususnya pada tahap pembersihan data untuk menjaga validitas medis, serta dalam tahap penanganan data *imbalanced* guna mengarahkan fokus model pada kasus-kasus minoritas yang relevan secara klinis. Pendekatan ini terbukti meningkatkan performa model ensemble, dengan potensi F1-Score mencapai 85%.

1.7 Kontribusi Penelitian

Berdasarkan eksplorasi hasil *literature review* yang sudah dilakukan, belum pernah ada penelitian yang mengusulkan pendekatan *preprocessing* dan model prediksi penyakit kronis dengan menerapkan *domain knowledge*. Sehingga, pada penelitian ini memiliki beberapa kontribusi ilmiah sebagai berikut:

1. Pendekatan *preprocessing* baru untuk menangani data yang *imbalanced* dalam rangka meningkatkan performansi prediksi kronis penyakit seseorang.
2. Model baru untuk melakukan prediksi terhadap seseorang terkena satu atau beberapa dari penyakit kronis.

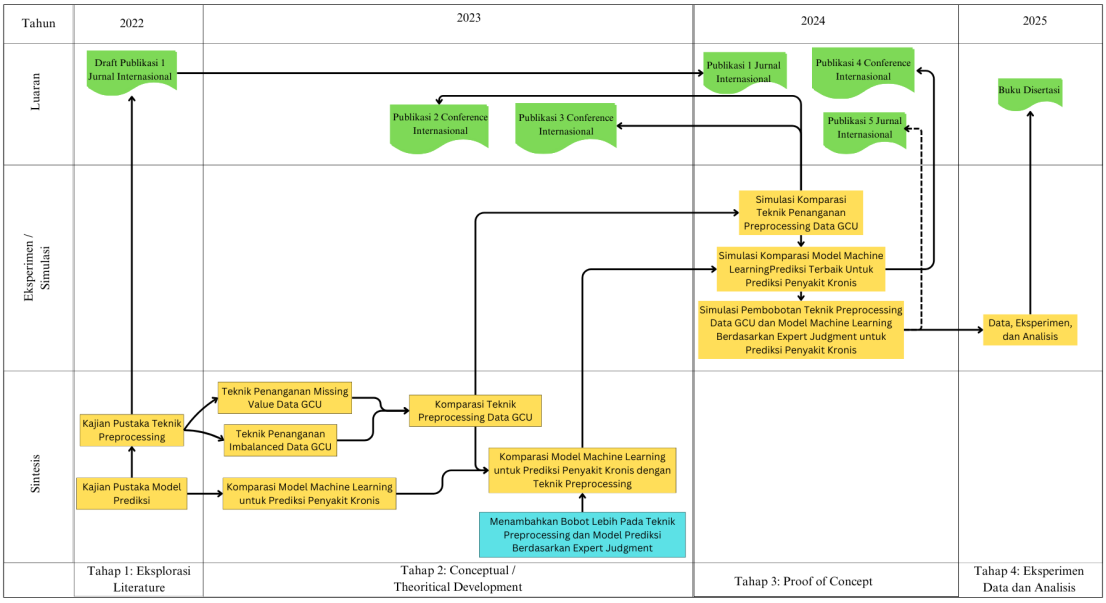
1.8 Batasan Masalah

Beberapa batasan masalah terhadap penelitian ini meliputi hal-hal berikut ini:

1. Data yang digunakan hasil pemeriksaan secara medis oleh individu (*general checkup*) dari Yakes Telkom mulai tahun 2019-2021.
2. Jenis penyakit kronis (tidak menular) yang digunakan ada yaitu diabetes mellitus, stroke, kanker, gagal ginjal, serangan jantung, dan hipertensi.
3. Peran *domain knowledge* dalam penelitian ini hanya sebatas memberikan pengetahuan fitur mana saja yang paling berpengaruh pada setiap penyakit.
4. Analisa hasil yang dilakukan hanya berfokus kepada penggunaan penerapan bobot pada algoritma *imbalanced* dan model *machine learning*.

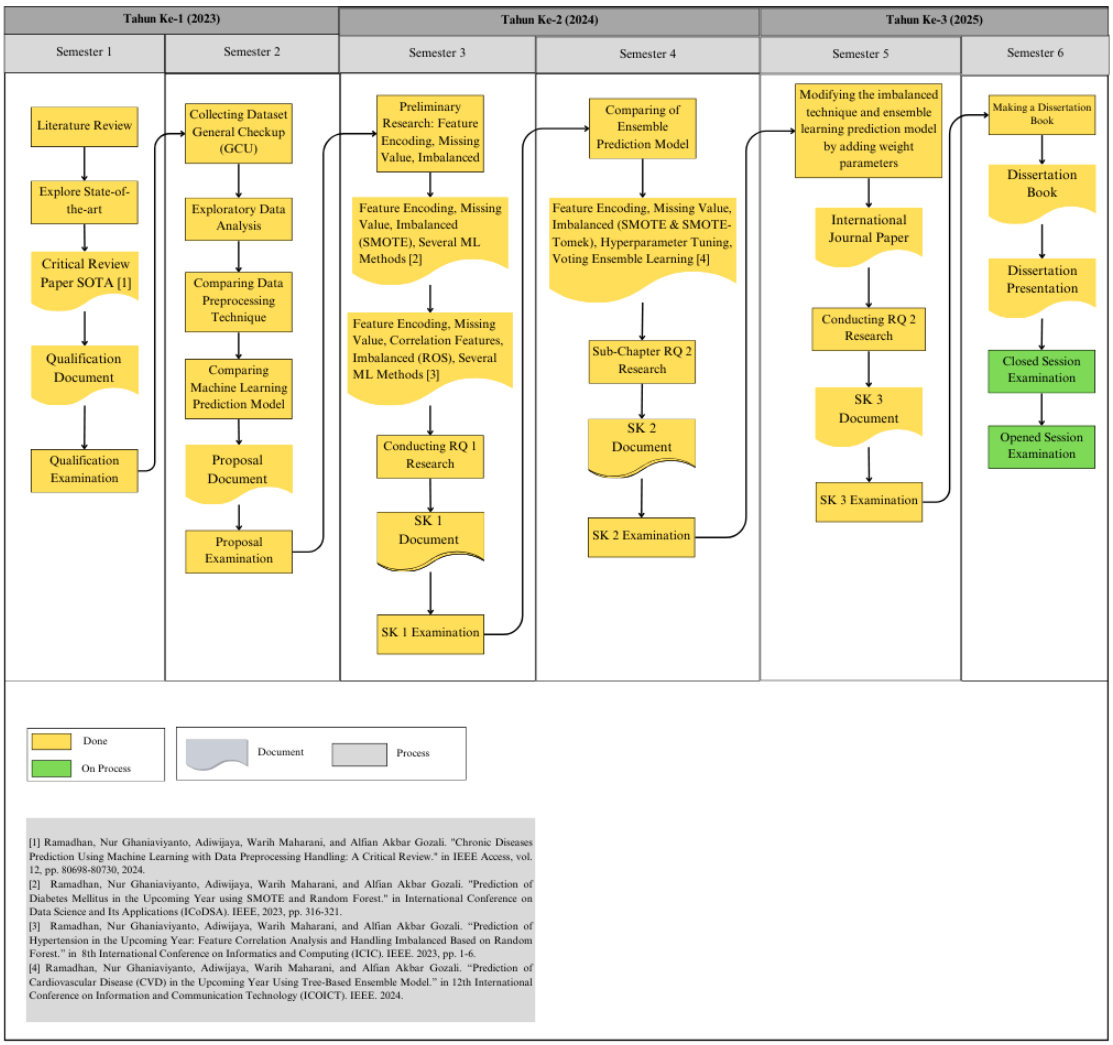
1.9 Peta Jalan Penelitian

Peta jalan untuk penelitian ini disusun dengan tujuan memastikan bahwa prosesnya akan mengikuti rencana awal yang telah ditentukan. Pada Gambar I.1 merupakan roadmap penelitian yang kami gunakan dan pada Gambar I.2 merupakan ringkasan cakupan kemajuan penelitian setiap semester:



Gambar I.1 Roadmap Penelitian

Pada tahap eksplorasi literatur telah dilakukan dengan mengumpulkan literatur-literatur dengan topik tentang prediksi penyakit kronis. Kajian Pustaka telah dilakukan dengan *mereview* lebih dari 150 literatur dan hasilnya telah *published* di jurnal IEEE Access (Q1). Pada tahap *conceptual / theoretical development* akan dibangun pendekatan dan model baru, yaitu pendekatan *preprocessing* berdasarkan *domain knowledge* dalam mendiagnosis penyakit kronis dan model *ensemble learning* untuk prediksi penyakit kronis. Mengingat penelitian tentang prediksi penyakit kronis menggunakan *domain knowledge* masih jarang maka diharapkan akan menjadi penelitian dan publikasi yang pertama. Seluruh model yang dibangun diimplementasikan ke dalam program komputer untuk dilakukan *proof-of-concept*. Sementara, untuk evaluasi hasil menggunakan *automatic evaluation* dan *manual expert evaluation*.



Gambar I.2 Ringkasan Cakupan Penelitian Setiap Semester

Pada Gambar I.2 merupakan ringkasan cakupan penelitian yang sudah dilakukan. Pada tahun pertama dimulai dengan melakukan *literature review* pada penelitian-penelitian prediksi penyakit kronis yang sudah ada saat ini. Hasil yang didapatkan dari *literature review* berupa *state of the art* yang akan menjadi gap penelitian disertasi. Selain itu, juga menghasilkan sebuah paper *Systematic Literature Review* (SLR) yang sudah *published* di jurnal internasional bereputasi Q1 (IEEE Access). Pada tahun pertama juga berhasil mendapatkan dataset general checkup (GCU) untuk penyakit kronis dengan rentang waktu datanya 2019-2021. Dataset tersebut dilakukan eksplorasi untuk di analisis karakteristik datanya. Selain itu, dataset dilakukan eksperimen dengan beberapa teknik *preprocessing* dan model *machine learning* yang sudah tersedia saat ini untuk melihat *best performed* teknik dan model yang digunakan.

Pada tahun kedua dilakukan eksperimen awal dengan menggunakan teknik preprocessing dan model *machine learning* dengan performa terbaik. Hasil dari eksperimen tersebut dihasilkan dua paper konferensi internasional (ICODSA dan ICIC), sekaligus menjawab pertanyaan RQ1. Pada tahun kedua juga telah dilakukan perbandingan terhadap model prediksi berbasis ensemble. Hasil perbandingan tersebut dituliskan dalam bentuk artikel konferensi internasional (ICOICT), hal tersebut merupakan bagian dari sub RQ2 yang akan dilakukan *embedd domain knowledge* pada tahun ketiga.

Pada tahun ketiga akan berfokus pada teknik SMOTE dan model ensemble dengan menambahkan parameter bobot. Parameter bobot yang dimaksud yaitu dengan memberikan bobot pada masing-masing fitur, dengan nilai bobotnya didapatkan dari formula matematika hasil integrasi dengan *domain knowledge* (dokter). Hasil dari menambahkan bobot tersebut akan dituliskan dalam bentuk artikel jurnal internasional bereputasi Q3 (IJICIC), sekaligus menjawab pertanyaan RQ2. Pada tahun ketiga juga dilakukan pembuatan buku disertasi guna mempersiapkan sidang disertasi secara tertutup dan terbuka.

1.10 Daftar Publikasi dan HKI Disertasi

Pada saat menempuh studi doktoral informatika di Universitas Telkom berhasil menghasilkan 5 publikasi internasional yang secara rinci dapat dilihat pada Tabel I-1. 3 publikasi di konferensi internasional terindeks IEEE dan scopus serta 2 publikasi di jurnal internasional dengan SJR Q1. Adapun status publikasi yang sudah ada adalah 4 artikel yang telah dipublikasikan (3 di konferensi internasional dan 1 di jurnal internasional SJR Q1), serta 2 artikel di jurnal internasional SJR Q3 yang sudah *accepted* dan menunggu *publish* bulan Agustus 2025.

Tabel I-1 Daftar publikasi yang berkaitan dengan disertasi prediksi penyakit kronis

No	Judul	Tahun	Nama Jurnal / Konferensi	Keterangan	Status
1	Prediction of Diabetes Mellitus in the Upcoming Year using SMOTE and Random Forest	2023	ICODSA	Konferensi internasional terindeks scopus	*Published*
2	Prediction of Hypertension in the Upcoming Year: Feature Correlation Analysis and Handling Imbalanced Based on Random Forest	2023	ICIC	Konferensi internasional terindeks scopus	*Published*

No	Judul	Tahun	Nama Jurnal / Konferensi	Keterangan	Status
3	Chronic Diseases Prediction Using Machine Learning with Data Preprocessing Handling: A Critical Review	2024	IEEE ACCESS	Jurnal Internasional SJR Q1	*Published*
4	Prediction of Cardiovascular Disease (CVD) in the Upcoming Year Using Tree-Based Ensemble Model	2024	ICOICT	Konferensi internasional terindeks scopus	*Published*
5	Analyzing Risk Factors and Handling Imbalanced Data for Predicting Stroke Risk Using Machine Learning	2025	IJAIN	Jurnal Internasional SJR Q3	*Published*
6	Modified SMOTE and Ensemble Learning Based on Expert Judgment for Chronic Diseases Prediction	2025	IJICIC	Jurnal Internasional SJR Q3	*Published*
7	Enhancing SMOTE Using Euclidean Weighting for Imbalanced Classification Dataset	2025	JADS	Jurnal Internasional SJR Q3	*Accepted*