Indonesian Journal of Cardiology
Indonesian J Cardiol 2023:44:158-165
pISSN: 2830-3105 / eISSN: 2964-7304
doi: 10.30701/ijc.1689

Review

The Use of Artificial Intelligence (AI) to Predict Heart Failure in Type II Diabetes Mellitus Patients: A Systematic Review

Lies Dina Liastuti[1,2,3], Averina Geffanie Suwana[2], Muhammad Aji Muharrom[2], Aruni Cahya Irfannadhira[2], Yosilia Nursakina[1,2].

Abstract

Introduction: Heart failure (HF) is a critical concern for individuals with Type II Diabetes Mellitus (T2DM), significantly increasing morbidity and mortality rates. Artificial Intelligence (AI) and machine learning hold promise in enhancing predictive capabilities and guiding personalised interventions. This systematic review evaluates existing AI models' effectiveness in predicting HF complications in T2DM patients.

Methods: A comprehensive literature search was performed across 8 different databases for studies investigating the effectivity of AI models on predicting HF complications in T2DM patients. Each literature was screened against the eligibility criteria. Subsequently, the included studies were critically appraised against the IJMEDI checklist for quality assessments of machine learning studies.

Results: A comprehensive literature search identified 8 relevant studies, predominantly from European, North American, and Southeastern populations. These studies utilised multi-centered registries and electronic medical records to develop AI models predominantly focused on supervised learning algorithms. While the AI models had promising performance, these models lack external validation and transparency in model development. Most of the studies were conducted in high-income countries, leaving a gap in applicability for low-to-middle income countries with differing demographics and risk factors. Hence hindering with study reproducibility and clinical applicability. Moreover, variations in outcome definitions and input features underscore the need for standardised approaches. Despite these limitations, AI models offer valuable insights into HF risk assessment in T2DM, highlighting the importance of further validation and reproducibility for clinical integration.

Conclusion: This review highlights the potential of machine learning models in predicting heart failure (HF) risk among patients with Type II Diabetes Mellitus (T2DM). The models demonstrated promising performance internally. Future research should focus on including a diversified population and external validation to ensure broader applicability and reliability in clinical practice.

(Indonesian J Cardiol. 2023;44:158-165)

Keywords: *Artificial Intelligence, Heart Failure, Type 2 Diabetes Mellitus, Machine Learning*.

[1] Dr. Cipto Mangunkusumo Hospital, Indonesia
[2] Faculty of Medicine, University of Indonesia, Indonesia
[3] Departemen of Cardiology and Vascular, National Cardiovascular Center Harapan Kita, Indonesia

Correspondence:
Lies Dina Liastuti,
Dr. Cipto Mangunkusumo Hospital, Indonesia, Faculty of Medicine, University of Indonesia, Indonesia, Departement of Cardiology and Vascular, National Cardiovascular Center Harapan Kita, Indonesia.
Email: liesdinaliastuti@gmail.com.

Introduction

Heart failure (HF) is a multifactorial condition characterized by inefficient myocardial performance to pump blood supply throughout the body, leading to various symptoms and complications. Among its numerous risk factors, Type II Diabetes Mellitus (T2DM) stands out as a significant contributor, significantly augmenting the likelihood of HF development. It has been found that subjects with type 2 diabetes have over twice the risk of incident heart failure than people without diabetes.[1] This can occur due to the major risk factors for heart failure are commonly present in type 2 diabetes patients such as hypertension, advanced age, obesity, sleep apnoea, dyslipidemia, and coronary heart disease (CHD).[2] Moreover, diabetes itself also independently contributes to the development and progression of heart failure.[3]

The survival and prognosis of heart failure patients with diabetes is worse approximately by half of that observed in non-diabetic patients. Heart failure is also the emerging leading cause of death in type 2 diabetes patients.[4-5] With the promising advancement of Artificial Intelligence (AI), particularly in machine learning, there has been a growing interest in leveraging these technologies to predict and manage heart failure in patients with T2DM.[6] Due to its ability to analyse multidimensional data, machine learning has the potential to increase predictive model performance. This can result in better prevention and introduce a more targeted intervention.

There are several studies regarding the use of machine learning to predict HF complications in type 2 diabetes, but none have been utilised and made it to clinical care guidelines. This systematic review aims to evaluate effectiveness, performance, applicability, and limitations of existing machine learning models in order to predict the risk, early detection, and make personalised management strategies for heart failure complications in T2DM patients.

Methods

A literature search was performed to study existing AI models in predicting heart failure incidences in people with Diabetes Mellitus type II (T2DM). We used the Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) Guidelines to report our findings.[1]

Search Strategy

A comprehensive literature search was conducted in multiple databases such as Cochrane CENTRAL, PubMed, Embase, ScienceDirect, Scopus, IEEE Xplore, Google Scholar, and Online Wiley database for publications published up to 14 March 2024. The keywords used for the search were "Cardiovascular disease", "CVD", "Type 2 diabetes", "prognosis", "predict outcome", "Prediction", "heart failure", "cardiac failure", "Machine learning" and "Artificial intelligence". The full search strategy was provided in the Supplementary **Table 1**.

Eligibility Criteria

The title and abstract of each literature were screened against the eligibility criteria which is listed below in **figure 1**. The inclusion criterias are studies that included adult patients (aged above 18 years old) with T2DM, predicted the risk of T2D developing heart failure, developed AI models, prediction models, and publications in English text. The exclusion criterias are as follows: 1) studies not written in English, 2) non-AI predictive models, 3) study such as review articles, commentaries, and conference reports.

Literature Selection

Literature search was performed on 8 databases which produced 1160 articles, and 472 duplicates were removed prior to screening. Based on the eligibility criteria, 32 studies from 104 screened articles were evaluated in full text to which 23 were excluded due to the following reasons: 1) AI was not involved, 2) different target population (not adults with type II diabetes mellitus patients), 3) not in English, 4) no predictive outcome, 5) Outcome did not involve heart failure, 6) no follow up. This yielded a total of 9 to be critically appraised **(Figure 1**).

Table 1. Table of Study characteristics.

Author	Year	Sample size	Country	Age (mean ± SD), median	Sex (% women)	BMI kg/m2 (mean ± SD)	Race/ ethnicity	Mean age at diagnosis, years	HbA1c (%) (mean ± SD)
Dworzynski	2020	203.517	Denmark	61.44	47.03	-	-	-	-
Abegaz	2023	9.059	United States	-	57.4	35.3 ± 7.2	Whites (46.6%)	-	-
Segar	2019	8.756	US & Canada	62.7 ± 6.6	38.5	32.1 ± 5.4	Black 18.5% Hispanic 7.5% Other 11.3% White 62.7%	-	8.3 ± 1.1
Basu	2017	9.635	US & Canada	62.8 ± 6.7	38	32.2 ± 5.4	Black 19% Hispanic/ Latino 7%	-	8.3 ± 1.1
Gandin	2023	10.614	Italy	72 ± 11	42	28.7 (23.4, 34)	-	-	HF free: 6.70 (6.24, 7.50) HF: 6.78 (6.20, 7.50)
Kanda	2022	217.054	Japan	67.7	42.2	-	-	-	7.09
Tee	2023	17.389	Singapore	-	Class 1: 51 Class 2: 48.3 Class 3: 45.8 Class 4: 42.4 Class 5: 46.5	-	Chinese Indian Malay Other races	Class 1: 61.6 ± 10.9 Class 2: 57.8 ± 10.7 Class 3: 49.2 ± 11.7 Class 4: 55.4 ± 10.5 Class 5: 53.2 ± 12.1	Class 1: 6.23 (0.59) Class 2: 7.16 (0.86) Class 3: 10.6 (2.14) Class 4: 7.69 (1.81) Class 5: 8.73 (1.7)
Mora	2023	490.762	Spain	79.23	45.67	-	Spanish (93.58%)	-	-

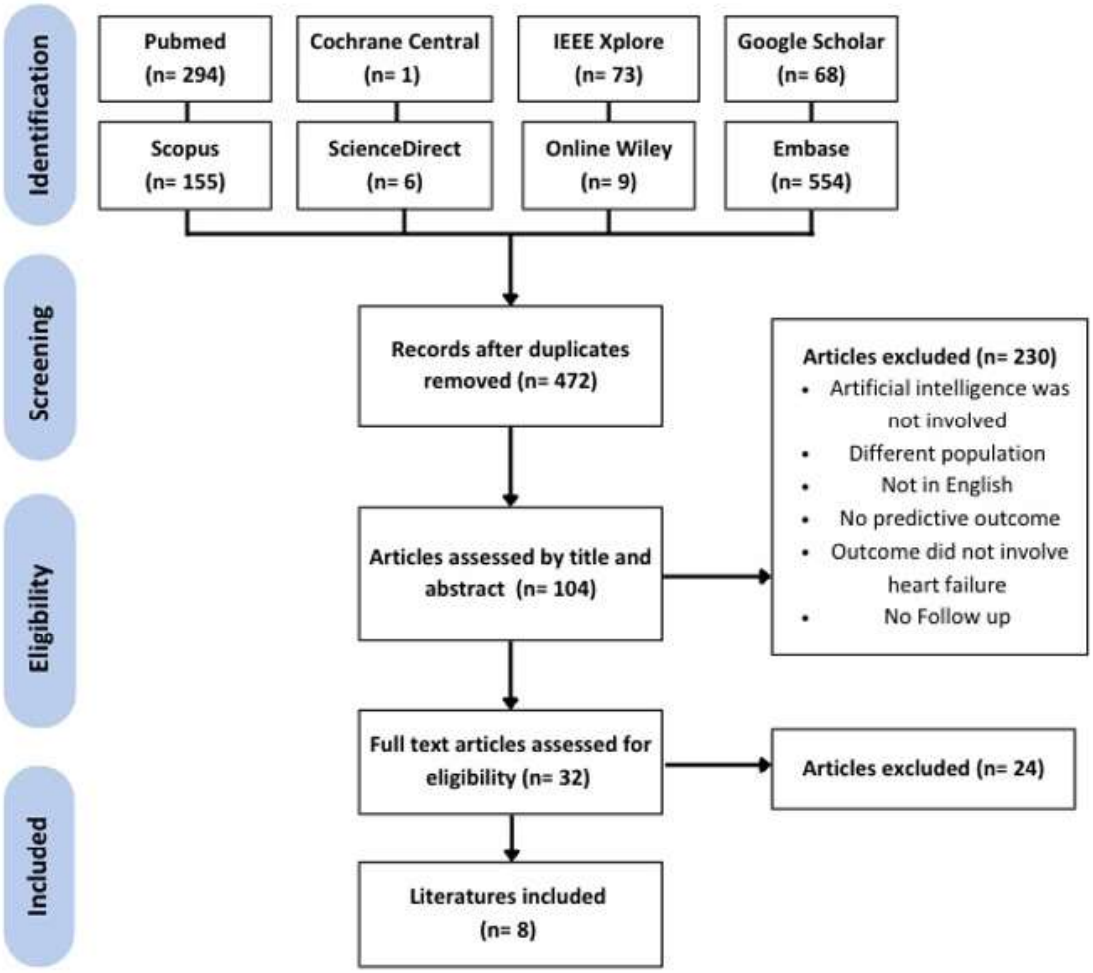

Figure 1. Flowchart of literature search

Results

Data Sources

All 8 studies have included patient demographics used to derive AI models for heart failure risk among people with type 2 diabetes as shown in **table 1**.[2–9] Derivation of the models for Segar et al[7] and Basu et al[3] were carried out on the same dataset. Dworynzki, Abegaz, and Tee have not included the mean age of patients included in the study.[2,8,9] Two studies did not mention the handling of missing data.[2,6] All studies mentioned subject selection criteria from the dataset.

Only 3 studies had applied external validation.[2,6] Segar et al[7] used a subset of ALLHAT participants. Basu used data from the Look AHEAD participants. Kanda et al[5] used external data from another population which is the Real World Data medical records. The rest of the included studies used internal validation strategy to evaluate the developed model.

Model Development

Table 2 presents a summary of model development from the selected studies. Multiple time frame outcomes were analyzed across the studies. Four studies analyzed the 5-year risk of heart failure. Three studies developed multiple models predicting heart failure for different time frames, ranging from 1-year to 5-year risk. While Basu et al[3] used a 10-year time period. Four studies conducted time-to-event analysis with proportional hazard regression models. One study by Abegaz et al[2] did not mention a specific time frame and used the presence

Table 3. Summary of the best performing Artificial intelligence models in each included studies for heart failure risk prediction in type 2 diabetic patients.

Author	Year	Internal validation setup			Model Outcome	Internal Performance
		Validation strategy	Data split / Cross-validation	Stratified data split or not		Area under receiver operating characteristics curve (AUROC)
Dworzynski	2020	Internal validation	Random 70-20-10 split	Yes	5-year HF risk	Gradient Boosting AUROC 0.80 (0.78-0.81)
Abegaz	2023	Cross-validation technique	80-20 split	N/A	Presence of MACE	XGBoost AUROC (0.80)
Segar	2019	Internal validation	Random 50-50 split, repeated 1,000 times	No	Cox PH + 5-year integer-based risk score	C-index (AUC analog) 0.77 (derivation dataset)
Basu	2017	Cross-validation technique	10-fold CV	N/A	Cox PH	C-index 0.75 (derivation dataset)
Gandin	2023	External validation	70-15-15 stratified HF-noHF split	Yes	Cox PH + PHNN + 2-year and 5-year risk	C-index 0.768
Kanda	2022	External validation	80-20 split	Yes	Cox PH + PHNN +	AUC of 1-year prediction: 0.728 AUC of 2-year prediction: 0.743 AUC of 3-year prediction: 0.740 AUC of 5-year prediction: 0.799
Tee	2023	Internal validation	80-20 split	Yes	HbA1c clusters + Cox PH	Bayesian Information Criterion 0.031 C-index 0.846
Mora	2023	External validation	80-20 split	yes	1, 2, 3, and 5 year risk of HF	AUC of 1-year prediction: 0.72 AUC of 2-year prediction: 0.70 AUC of 3-year prediction: 0.69

Table 4. Reporting quality of each included studies based on the IJMEDI checklist.

Author	Year	Problem Understanding (10)	Data Understanding (6)	Data Preparation (8)	Modeling (6)	Validation (12)	Deployment (8)	Total (50)
Abegaz(2)	2023	8	2	5	6	8	1	30
Basu(3)	2017	6.5	4	3	6	7	6.5	33
Gandin(4)	2023	8	3	4	6	8	3	32
Kanda(5)	2022	9	3	8	6	8	3.5	37.5
Mora (6)	2023	8	4	2	6	9	4	33
Segar(7)	2019	6.5	5	3	6	8	2.5	31
Tee(8)	2023	9	4	2	6	5.5	7	33.5
Dworzynski(9)	2020	10	5	2	6	6.5	1.5	31

of heart failure from the database as the outcome.

All studies implemented data-driven machine learning approaches for development of the predictive models. Seven out of the 8 studies were focused on supervised learning algorithms, except for the study by Tee et al[8] that utilised an unsupervised learning algorithm that is able to cluster their subjects to classes that correspond to significantly different risks of developing heart failure. Multiple studies evaluated performance of different machine learning algorithms. Three studies showed gradient-boosted trees as their best-performing algorithm. Deep neural network, logistic regression, random survival forest for feature selection followed by proportional hazard, and proportional hazard regression were the best-performing algorithms in the rest of the studies.

For model validation, 6 out of 8 studies performed data splitting with a held-out dataset for the validation, with 80-20 split being the most common split ratio. Basu et al[3] performed 10-fold cross-validation while Segar et al[7] evaluated the model with a 50-50 split that is repeated 1000 times. All studies experienced challenges pertaining to an imbalanced dataset, but only three studies reported a strategy to address it.

Reproducibility has not been well-addressed among the developed predictive models. While Segar et al[7] and Basu et al[3] provided an online calculator for their models, calculations relied on the server and no source code was provided. No other studies provided source code or data for their developed models, which render it impossible to independently cross-validate the developed models on external data sources.

A summary of input features for each of the developed models was also included in **Table 2**. Among the input features, age was the only common feature across all the studies. With respect to the interpretability of the models, Mora et al[6] and Kanda et al[5] carried out SHAP (SHapley Additive exPlanation) analysis to determine the importance of each feature included in the model. Both commonly revealed age as one of the most important features. Abegaz et al[2] conducted a stepwise forward selection for developing the deep neural network model as well as calculating partial dependence analysis on each of the features, reporting the partial dependence plot. Dworzynski et al.[9] analysed feature importance using the built-in method for gradient boosted model.

Model Performance

Table 3. Summary of the best performing Artificial intelligence models in each included studies for heart failure risk prediction in type 2 diabetic patients

Multiple definitions of heart failure outcome were used in the studies. Some only considered incident hospitalisation or death due to heart failure while others included first encounter of inpatient and outpatient diagnosis of heart failure. A summary is available on **Table 3**.

The area under the receiver operating characteristics (AUROC) was the most common metrics used for evaluating a model's discriminative performance in all of the selected studies. Three studies which developed a proportional hazards regression model reported the concordance statistics (C-index), which can be considered an analog of AUROC for proportional hazards analysis. **Table 3** summarised the performance of the best performing models within each study. AUROC ranges from 0.69 to 0.80 while C-index ranges from 0.75 to 0.85. The results from other metrics were included based on the reports from each study.

Reporting Quality

Table 4 summarises the quality report based on the IJMEDI checklist for assessment of medical AI model development and validation. 7 out of the 8 studies fell had medium quality (20 - 34.5 points) based on the checklist. While Kanda et al[5], with a score of 37.5 leading to a high quality study. While all studies had a relatively low score on data preparation, Kanda et al[5] reported a complete and comprehensive report on data preparation with a perfect 8 out of 8 score

Discussion

This review included 8 articles on AI models for heart failure risk prediction specifically in type 2 diabetes patients. All the models were developed in the European, North American and Southeast Asia population. While it includes a range of varying ethnicities, there is a lack of models developed for populations of low-to-middle income countries that have different risk factors, comorbidities, and lifestyle which greatly impact the prevalence of diabetes and HF risk. The average age

included in the studies is 60-70 years. The mean age found in these studies illustrates that older age is prone to higher risk of diabetes and HF. There are 4 studies which included the calculation of mean BMI, three of which had a mean of BMI higher than 30 reflecting mostly obese patients in the diabetic population.

All of the studies used multi-centered registries and electronic medical records to develop and validate the AI models. Most of the studies included an imbalance dataset however only three had applied treatment to address them. While this helps represent the real prevalence in the population, there is a risk that the prediction will be biased towards larger subgroups. All the studies used AUROC in model evaluation to assess the AI model performance. Some studies also included sensitivity, specificity, accuracy, and precision metrics. Most of the studies developed up to 4 different AI models and compared their performance. All of the studies had promising results from the AI models with an AUROC ranging from 0.72 to 0.8.

All of the AI models achieved good performance internally and externally. However, only 3 had external validation but none of which used a population with a different demographics from the development data, which may limit their generalisability. Moreover, they are also not validated with prospective data; therefore, vulnerability to real-world challenges such as data drifting is not yet proven. Additionally, real-world deployment of these machine learning models may be challenging, especially for complex models that need hundreds of data points. Furthermore, no studies published online models or elsewhere. Studies such as Segar and Basu published an online calculator to be used for HF prediction in type II DM patients, however the model codes are not published and therefore limit efforts to replicate or externally validate the models. Moreover, the generalizability of the AI models might be lacking for low-to-middle-income countries due to limited study scope.

Additionally, there is a wide array of outcomes including major adverse cardiovascular events (MACE), hospitalisation, length of stay, chronic kidney disease, and inpatient and outpatient diagnosis of HF. Aside from comorbidities affecting heart failure risk, medical procedures, medications also play an essential role in determining HF risk in diabetes patients. 4 of the studies have included them in developing the prediction models, thereby more closely reflecting the general population with diverse medical backgrounds. Furthermore, there are 4 studies that studies the risk of heart failure in 1,2,3-, 5-, and 10-years period. As time is an important factor for HF prediction risk, AI models could be an important tool in guiding clinicians on medical decisions for their patients.

Included studies revealed significant variability in model outcomes and performance of the machine learning models. This variability arises from the different outcome definitions across the evaluation metrics, and model development techniques. Each of the studies employed varied definitions of heart failure outcomes, ranging from hospitalization and death to initial diagnosis in both inpatient and outpatient settings. These differences might impact the generalizability of the findings. All studies used AUROC to evaluate the model performance reducing the difference in evaluating model performance. Lastly, each study had employed their own choice of algorithms. This underscores the challenge in generalizing the review's findings.

Future research should focus on standardizing outcome definitions, improving validation strategies, and ensuring transparency in model development. Additionally, including diverse populations, especially from low-to-middle-income countries, will improve the models' generalizability and relevance across different demographic and clinical settings.

Conclusion

In conclusion, we have identified 8 studies that have developed AI risk prediction models for heart failure in people with type 2 diabetes. All the AI models have good performance with high sensitivity and specificity to predict HF. Future studies should focus on including different populations to ensure higher applicability and reliability in clinical practice..

References

1. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Syst Rev. 2021 Mar 29;10:89.
2. Abegaz TM, Baljoon A, Kilanko O, Sherbeny F,

Ali AA. Machine learning algorithms to predict major adverse cardiovascular events in patients with diabetes. Comput Biol Med. 2023 Sep;164:107289.
3. Basu S, Sussman JB, Berkowitz SA, Hayward RA, Yudkin JS. Development and validation of Risk Equations for Complications Of type 2 Diabetes (RECODe) using individual participant data from randomised trials. Lancet Diabetes Endocrinol. 2017 Oct;5(10):788–98.
4. Gandin I, Saccani S, Coser A, Scagnetto A, Cappelletto C, Candido R, et al. Deep-learning-based prognostic modeling for incident heart failure in patients with diabetes using electronic health records: A retrospective cohort study. Cannatà A, editor. PLOS ONE. 2023 Feb 21;18(2):e0281878.
5. Kanda E, Suzuki A, Makino M, Tsubota H, Kanemata S, Shirakawa K, et al. Machine learning models for prediction of HF and CKD development in early-stage type 2 diabetes patients. Sci Rep. 2022 Nov 21;12(1):20012.
6. Mora T, Roche D, Rodríguez-Sánchez B. Predicting the onset of diabetes-related complications after a diabetes diagnosis with machine learning algorithms. Diabetes Res Clin Pract. 2023 Oct;204:110910.
7. Segar MW, Vaduganathan M, Patel KV, McGuire DK, Butler J, Fonarow GC, et al. Machine Learning to Predict the Risk of Incident Heart Failure Hospitalization Among Patients With Diabetes: The WATCH-DM Risk Score. Diabetes Care. 2019 Dec 1;42(12):2298–306.
8. Tee C, Xu H, Fu X, Cui D, Jafar TH, Bee YM. Longitudinal HbA1c trajectory modelling reveals the association of HbA1c and risk of hospitalization for heart failure for patients with type 2 diabetes mellitus. Lang CC, editor. PLOS ONE. 2023 Jan 20;18(1):e0275610.
9. Dworzynski P, Aasbrenn M, Rostgaard K, Melbye M, Gerds TA, Hjalgrim H, et al. Nationwide prediction of type 2 diabetes comorbidities. Sci Rep. 2020 Feb 4;10(1):1776.