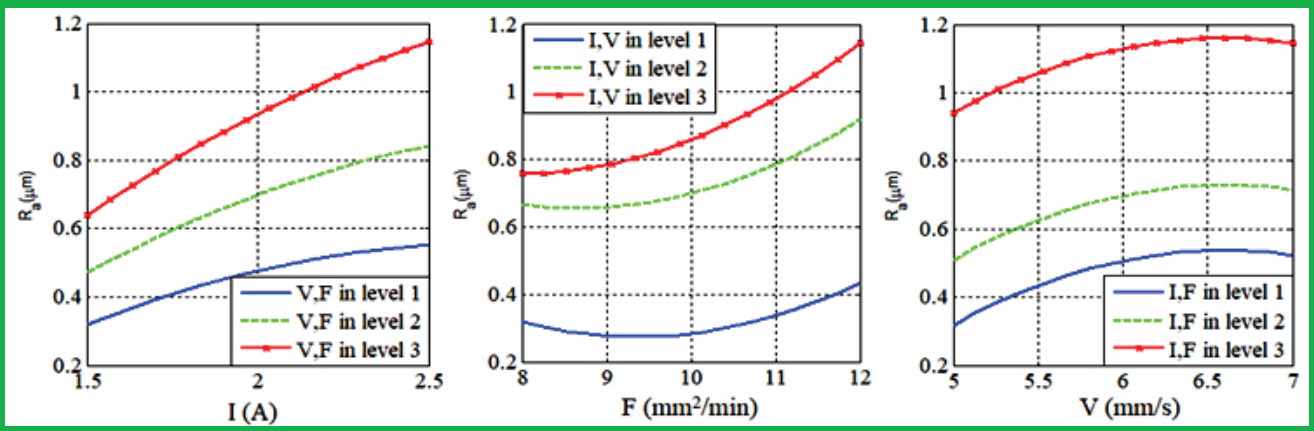
p-ISSN 3030-4830

Tạp chí
KHOA HỌC VÀ CÔNG NGHỆ ỨNG DỤNG

JOURNAL OF APPLIED SCIENCE AND TECHNOLOGY

Số 48
Tháng 12/2025



TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT HƯNG YÊN
HUNG YEN UNIVERSITY OF TECHNOLOGY AND EDUCATION

MỤC LỤC

p-ISSN 3030-4830

p-ISSN 3030-4830

SMOTE-ENSEMBLE: TỔNG QUAN KỸ THUẬT CÂN BẰNG DỮ LIỆU VÀ MÔ HÌNH HỌC MÁY KẾT HỢP TRONG DỰ ĐOÁN SỚM BIẾN CHỨNG VÕNG MẠC ĐÁI THÁO ĐƯỜNG

Hà Ngọc Tuân, Phạm Thị Ánh Hương, Trần Thị Thu Huyền, Ngô Thị Lan Anh*
Trường Đại học Sư phạm Kỹ thuật Hưng Yên
** Tác giả liên hệ: ntlananh.utehy@gmail.com*

Ngày tòa soạn nhận được bài báo: 15/07/2025
Ngày phản biện đánh giá và sửa chữa: 19/09/2025
Ngày bài báo được duyệt đăng: 08/12/2025

Tóm tắt:

Bệnh lý thoái hóa võng mạc xuất phát từ rối loạn đường huyết kéo dài (Diabetic Retinopathy - DR) hiện đang là mối đe dọa hàng đầu đối với khả năng nhìn của con người trên phạm vi toàn thế giới. Điểm đặc thù của căn bệnh này nằm ở chỗ các tổn thương ban đầu diễn ra hoàn toàn âm thầm, khiến người bệnh không hề hay biết cho đến khi mức độ hủy hoại đã vượt ngưỡng có thể cứu vãn. Khi triển khai các hệ thống rà soát hình ảnh đáy mắt trên diện rộng, một rào cản kỹ thuật nổi bật xuất hiện: trong mọi tập dữ liệu thu thập được, những trường hợp bệnh nặng đòi hỏi xử lý cấp bách luôn chiếm số lượng cực kỳ khiêm tốn so với các mẫu bình thường hoặc nhẹ. Công trình này giới thiệu một chiến lược phương pháp luận kết hợp giữa kỹ thuật tạo mẫu tổng hợp SMOTE và kiến trúc học máy tập thể (SMOTE-Ensemble) phục vụ mục đích thiết lập cơ chế nhận diện nguy cơ từ sớm. Trước tiên, chúng tôi mổ xẻ bản chất sinh lý bệnh học của DR song song với việc khảo sát cấu trúc phân bố dữ liệu, qua đó làm sáng tỏ ảnh hưởng bất lợi của hiện tượng chênh lệch tỷ lệ giữa các nhóm đối với khả năng phát hiện đúng của bộ phân loại. Kế tiếp, bài viết phân tích sâu thuật toán SMOTE gốc cùng các phiên bản nâng cấp như Borderline-SMOTE, ADASYN, Geometric SMOTE - những công cụ cho phép sản sinh ra các điểm dữ liệu nhân tạo một cách có chủ đích [1]. Ngoài ra, các chiến lược học tập thể bao gồm Bagging, Boosting, Voting, Stacking cũng được đặt dưới góc nhìn phân tích, đặc biệt nhấn mạnh vào SMOTEBoost và những cải tiến mới nhất. Bằng việc tổng hợp hơn 30 nghiên cứu đã công bố, chúng tôi chứng minh rằng phương thức SMOTE-Ensemble đem lại mức tăng từ 5 tới 18 điểm phần trăm trên các thang đo AUC, F1-score, recall cho phân nhóm bệnh nghiêm trọng [2]. Phần kết của nghiên cứu đề cập những khiếm khuyết còn tồn đọng, đồng thời vạch ra các con đường nghiên cứu tương lai như tích hợp mạng đối kháng sinh (GAN) nhằm làm giàu dữ liệu hình ảnh, học đa nhiệm vụ, hay xây dựng các kiến trúc mô hình có năng lực giải trình cao phục vụ việc đưa công nghệ vào ứng dụng thực tế tại cơ sở y tế.

Từ khóa: *SMOTE-Ensemble, học tập thể, Borderline-SMOTE, ADASYN, Geometric SMOTE, dữ liệu phân bố lệch.*

1. Đặt vấn đề

Trong số các di chứng thường gặp của tình trạng rối loạn chuyển hóa glucose kéo dài, thoái hóa cấu trúc mạch máu nhỏ tại võng mạc nổi lên như một vấn đề y khoa có tầm quan trọng đặc biệt. Khi bỏ qua giai đoạn vàng để phát hiện và xử trí, người bệnh đứng trước nguy cơ cao mất đi vĩnh viễn khả năng nhận thức ánh sáng. Dữ liệu do các tổ chức y tế quốc tế công bố chỉ ra rằng trong cộng đồng người mang bệnh đường huyết, tần suất xuất hiện tổn thương đáy mắt nằm trong biên độ 20–40%, với sự dao động phụ thuộc vào đặc điểm địa lý và cơ cấu dân cư. Nếu quy ra con số tuyệt đối trên bình diện toàn cầu, xấp xỉ 34,6% cá nhân mắc chứng rối loạn glucose huyết đã mang trong mình những dấu vết bệnh lý tại lớp tế bào nhận cảm ánh sáng – nói cách khác, cứ ba bệnh nhân tiểu đường thì nhiều hơn một người đã bước vào quá trình thoái hóa võng mạc. Đáng lo ngại hơn, có tới 7% trong số đó tiến triển sang thể tăng sinh – giai đoạn mà xác suất đánh mất thị giác tăng vọt. Một đặc trưng căn bản của bệnh lý này chính là diễn biến không triệu chứng; đại đa số cá nhân bị ảnh hưởng hoàn toàn không cảm nhận được sự bất thường nào cho tới khi mô võng mạc đã chịu tổn hại nặng nề và không còn cơ hội hồi phục toàn phần dù áp dụng mọi biện pháp điều trị hiện có. Bối cảnh đó đặt ra yêu cầu cấp thiết về việc kiến tạo những công cụ dự báo và phát hiện ở giai đoạn manh nha nhằm gìn giữ chức năng thị giác cho những ai đang chung sống với bệnh đường huyết.

Bức tranh dịch tễ học về rối loạn glucose máu đang biến chuyển theo chiều hướng đáng báo động trên quy mô hành tinh, kéo theo đó là sự bùng nổ số ca biến chứng mắt và tạo ra gánh nặng khổng lồ cho ngành y. Con số thống kê năm 2021 cho thấy lượng người trưởng thành sống chung với tình trạng đường huyết cao đã chạm ngưỡng 537 triệu – chiếm 10,5% tổng dân số trái đất và vượt xa hơn ba lần so với hai thập kỷ trước [3]. Theo các kịch bản dự phóng, đến năm 2045 con số này có thể leo thang

tới 783 triệu [4]. Tại Việt Nam, tỷ suất mắc bệnh đường huyết đã nhảy từ 2,7% vào năm 2002 lên quanh mức 7,3% tính đến năm 2020. Đi kèm với xu thế đó, tần suất gặp phải các rắc rối về mắt cũng leo thang: các khảo sát trong nước ước lượng rằng 20–35% người bệnh tiểu đường Việt Nam đã xuất hiện dấu hiệu thoái hóa võng mạc. Thực tế, bệnh lý mạch máu nhỏ tại đáy mắt do đường huyết cao hiện đứng đầu danh sách nguyên nhân gây suy giảm thị lực trong nhóm người ở độ tuổi làm việc. Hai nhân tố quyết định tới sự hình thành và tiến triển của biến chứng này là khoảng thời gian chung sống với bệnh cùng chất lượng kiểm soát nồng độ glucose. Dẫn chứng minh họa: trong nhóm bệnh nhân mắc thể phụ thuộc insulin (típ 1), sau 5 năm khoảng 25% đã có dấu hiệu tổn thương mắt, tỷ lệ này vọt lên 60% sau một thập kỷ và chạm ngưỡng 80% ở mốc 15 năm. Với thể không phụ thuộc insulin (típ 2), xấp xỉ 40% phát sinh biến chứng võng mạc sau 5 năm từ thời điểm chẩn đoán. Nghiên cứu nổi tiếng UKPDS đã định lượng rằng mỗi điểm phần trăm HbA1c được hạ xuống tương đương với việc cắt giảm khoảng 35% rủi ro gặp các tổn thương vi mạch, bao hàm cả thoái hóa võng mạc.

Căn cứ vào tình hình thực tiễn nêu trên, các cơ quan y tế hàng đầu đồng loạt đưa ra hướng dẫn yêu cầu mọi cá nhân mang bệnh đường huyết cần trải qua quy trình kiểm tra cấu trúc đáy mắt mỗi năm một lần nhằm kịp thời phát hiện bất kỳ dấu hiệu suy thoái nào. Dù vậy, trong thực hành, tỷ lệ tuân thủ khuyến cáo này còn nhiều thiếu sót, nhất là tại những vùng miền thiếu thốn đội ngũ bác sĩ chuyên về mắt hoặc hạ tầng khám chữa bệnh còn sơ khai. Hậu quả là không ít người chỉ biết mình mắc bệnh khi thị lực đã tụt dốc không thể vãn hồi. Đối diện với thực trạng đó, giới học thuật đang tập trung nguồn lực vào các phương án vận dụng máy tính thông minh và kỹ thuật học từ dữ liệu để sàng lọc những cá nhân tiềm ẩn nguy cơ cao bị thoái hóa võng mạc. Các thuật toán phân tích dữ liệu có thể khai thác kho thông tin lâm sàng dồi dào từ hồ sơ điện tử, các phép đo sinh hóa, các chỉ dấu sinh học để xác định những ca có dấu hiệu tiềm tàng hoặc dự báo xác suất hình thành biến chứng trong tương lai. Với lối tiếp cận này, cơ sở y tế có khả năng chủ động liên hệ những bệnh nhân thuộc nhóm rủi ro cao đến khám mắt trước khi họ xuất hiện bất kỳ triệu chứng nào. Một nghiên cứu tại Mỹ tiến hành trên 40.631 người bệnh tiểu đường đã xác nhận mô hình mạng nơ-ron nhiều lớp, khi được cho ăn dữ liệu hồ sơ điện tử kết hợp kỹ thuật cân bằng mẫu, có thể đạt AUC ≈ 0,80, với khả năng phát hiện đúng 72,2% và loại trừ đúng 74,2% khi nhận dạng những ca đã có tổn thương mắt nhưng chưa được chẩn đoán [5]. Kết quả này củng cố niềm tin vào tiềm năng của trí tuệ máy trong việc hỗ trợ sàng lọc biến chứng mắt ở quy mô lớn.

Bên cạnh những triển vọng đầy hứa hẹn, quá trình xây dựng công cụ dự báo biến chứng mắt từ dữ liệu bệnh nhân tiểu đường vẫn vấp phải không ít chướng ngại. Nổi bật nhất trong số đó là tình trạng bất đối xứng trầm trọng về tỷ lệ giữa các nhóm (class imbalance). Xét một tập dữ liệu y khoa điển hình, lượng hồ sơ thuộc nhóm đã phát sinh biến chứng thường thua xa nhóm chưa gặp vấn đề gì. Lấy ví dụ cụ thể, trong cộng đồng bệnh nhân tiểu đường, số người đã có thoái hóa võng mạc (chiếm khoảng 20–30%) thấp hơn đáng kể so với số chưa biến chứng (70–80%). Khoảng cách quá lớn này khiến hầu hết thuật toán học từ dữ liệu có khuynh hướng "ngả" về phía nhóm chiếm ưu thế (tức nhóm "chưa biến chứng"). Lý do là các thuật toán được thiết kế để tối đa hóa tỷ lệ phán đoán đúng tổng thể, do đó chúng có xu hướng phớt lờ hoặc gán nhãn sai những ca thuộc nhóm ít hơn (nhóm «đã biến chứng») bởi nhóm này chỉ chiếm phần nhỏ trong tổng mẫu. Kết cục là công cụ dự báo có nguy cơ «bỏ sót» đúng những bệnh nhân đang cần được chú ý nhất, kéo theo tỷ lệ phát hiện đúng (sensitivity/ recall) đối với nhóm biến chứng xuống thấp đáng lo ngại – điều hoàn toàn không thể chấp nhận trong bối cảnh sàng lọc y khoa, khi việc không nhận ra một ca bệnh thực sự có thể dẫn tới hậu quả khôn lường. Như Salmi và nhóm cộng sự đã đúc kết qua hơn mười năm nghiên cứu về dữ liệu y tế bất đối xứng, việc gán nhãn sai những mẫu thuộc nhóm thiểu số tiềm ẩn nguy cơ gây tổn hại nghiêm trọng cho sức khỏe người bệnh. Do vậy, việc xử lý triệt để bài toán mất cân bằng dữ liệu trở thành điều kiện tiên quyết để kiến tạo công cụ dự báo biến chứng mắt đạt hiệu năng mong muốn.

2. Dữ liệu và Phương pháp

2.1. Rào cản từ dữ liệu phân bố lệch

Đến thời điểm hiện tại, cộng đồng khoa học đã đề xuất nhiều hướng giải quyết cho vấn đề dữ liệu phân bố không đồng đều, có thể quy về một số trường phái chính: kỹ thuật điều chỉnh cấu trúc mẫu (resampling), phương pháp học có ý thức về chi phí (cost-sensitive learning), và các thuật toán được định hướng chuyên biệt cho dữ liệu lệch. Trong đó, kỹ thuật điều chỉnh cấu trúc mẫu chiếm vị trí phổ cập nhất bởi tính trực giác và thành tích đã được minh chứng qua hàng loạt thí nghiệm. Nhánh kỹ thuật này tách thành hai hướng cơ bản: tăng mẫu (oversampling) – tức bơm thêm dữ liệu cho phía ít hơn, và giảm mẫu (undersampling) – tức loại bớt dữ liệu từ phía nhiều hơn. Hướng giảm mẫu thực hiện việc cắt bỏ một phần dữ liệu từ nhóm áp đảo để thiết lập thế cân bằng với nhóm yếu thế, song hạn chế rõ ràng là làm mất đi khối lượng thông tin tiềm tàng đáng kể. Ở chiều ngược lại, hướng tăng mẫu tạo sinh các điểm dữ liệu mới cho nhóm yếu thế, cho phép tận dụng trọn vẹn nguồn dữ liệu hiện hữu đồng thời củng cố năng lực học của mô hình

đối với những trường hợp ít gặp. Một kỹ thuật tăng mẫu kinh điển và đặc biệt thích hợp với bối cảnh y sinh là SMOTE (Synthetic Minority Over-sampling Technique.

2.2. Kỹ thuật SMOTE – Tạo sinh điểm dữ liệu nhân tạo cho nhóm yếu thế

Kỹ thuật SMOTE lần đầu xuất hiện năm 2002 trong công trình của Chawla và đồng nghiệp, đưa ra cơ chế sản sinh mẫu nhân tạo dành riêng cho nhóm yếu thế với đích đến là tái thiết lập thế cân bằng cho tập huấn luyện. Khác hẳn với cách làm đơn giản nhất là sao chép ngẫu nhiên các mẫu yếu thế sẵn có, SMOTE tạo ra những điểm dữ liệu hoàn toàn mới bằng phép nội suy tuyến tính giữa các mẫu yếu thế nằm kề nhau trong không gian đặc trưng. Chi tiết quy trình như sau: với từng mẫu thuộc nhóm yếu thế, thuật toán xác định k người hàng xóm gần nhất (thông thường k = 5) cùng thuộc nhóm đó; kế đó chọn ngẫu nhiên một hoặc vài hàng xóm rồi phát sinh điểm mới trên đường nối điểm gốc với hàng xóm được chọn. Những điểm mới sinh ra mang tính chất tương đồng với dữ liệu yếu thế ban đầu, góp phần "nới rộng" phạm vi không gian đặc trưng của nhóm này. Sau khi áp dụng SMOTE, tập dữ liệu trở nên cân đối hơn về tỷ lệ giữa các nhóm; mô hình huấn luyện trên đó sẽ bớt đi xu hướng thiên vị và nâng cao năng lực nhận dạng nhóm yếu thế.

Hàng loạt kết quả thực nghiệm đã xác nhận SMOTE đem lại bước cải thiện rõ rệt về tỷ lệ phát hiện đúng cùng các chỉ số đánh giá phân loại đối với nhóm ít gặp. Nghiên cứu khởi nguồn của Chawla (2002) phối hợp SMOTE với bộ phân loại cây quyết định C4.5, ghi nhận khả năng nhận dạng chính xác nhóm yếu thế tăng vọt và chỉ số F-value tổng thể được cải thiện đáng kể so với kịch bản không dùng SMOTE. Trong địa hạt y sinh, SMOTE tỏ ra đặc biệt hữu dụng bởi dữ liệu bệnh lý vốn khan hiếm về số lượng và mang đặc điểm lệch nghiêm trọng. Salmi cùng cộng sự (2024) khi đánh giá hàng loạt nghiên cứu y sinh đã kết luận rằng các kỹ thuật như SMOTE có năng lực bổ khuyết hiệu quả cho các tập dữ liệu hạn chế, từ đó nâng tầm hiệu suất dự báo bệnh. Chẳng hạn, trong bài toán nhận diện sớm bệnh đường huyết, việc vận dụng SMOTE để tái cân bằng tập huấn luyện đã tiếp sức cho mô hình phát hiện tốt hơn những ca bệnh ở giai đoạn khởi phát (nhóm chiếm tỷ trọng thấp). Tương tự, với biến chứng mắt do tiểu đường, hoàn toàn có cơ sở để kỳ vọng rằng việc làm giàu dữ liệu cho nhóm bệnh nhân đã biến chứng sẽ giúp thuật toán học máy nắm bắt các đặc trưng của biến chứng một cách rõ nét hơn, thay vì bị "chìm đuối" trong số đông bệnh nhân chưa biến chứng.

Tuy vậy, hướng tăng mẫu nói chung và SMOTE nói riêng đòi hỏi phải được triển khai một cách thận trọng để phát huy tối đa hiệu năng. Nếu sử dụng quá liều, mô hình có nguy cơ rơi vào trạng thái khớp quá (overfit) với các mẫu nhân tạo hoặc hình thành đường phân chia không tối ưu. Bởi vậy, các công trình gần đây thường kết hợp SMOTE với các kỹ thuật phụ trợ nhằm củng cố độ chính xác và khả năng tổng quát hóa của mô hình. Nhiều phiên bản mở rộng của SMOTE cũng ra đời, tiêu biểu như Borderline-SMOTE, ADASYN (Adaptive Synthetic Sampling) và Geometric SMOTE, nhắm tới mục tiêu nâng chất lượng mẫu tổng hợp trong các tình huống dữ liệu phức tạp [1]. Cụ thể, Borderline-SMOTE chỉ tập trung phát sinh mẫu quanh những điểm yếu thế nằm sát ranh giới phân chia giữa các nhóm, còn ADASYN tự động điều tiết số lượng mẫu cần sinh tùy theo mức độ khó phân loại của từng điểm yếu thế – điểm nào «khó» hơn sẽ được sinh nhiều mẫu hơn [1]. Geometric SMOTE mở rộng cơ chế SMOTE gốc bằng cách phát sinh điểm mới trong các không gian hình học linh hoạt, có thể phát huy tác dụng với dữ liệu nhiều chiều [5]. Nhìn chung, các phiên bản mở rộng này giúp hạn chế nguy cơ sinh ra mẫu vô ích hoặc gây khớp quá, đồng thời dồn nguồn lực vào những trường hợp yếu thế khó học nhất [6].

2.3. Kiến trúc học máy tập thể trong dự báo sớm

Học tập thể (Ensemble learning) là chiến lược xây dựng bộ dự báo thông qua việc hợp nhất nhiều bộ dự báo thành phần nhằm khai thác "sức mạnh tập thể" của chúng. Nguyên lý "trí khôn đám đông" – rằng khi trung bình hóa ý kiến của nhiều cá thể thường thu được phán đoán tốt hơn – cũng phát huy hiệu quả trong lĩnh vực học máy: một nhóm nhiều bộ dự báo "yếu" có khả năng tạo ra kết quả chính xác và ổn định hơn so với một bộ dự báo "mạnh" đơn lẻ. Các chiến lược tập thể phổ biến bao gồm: Bagging (huấn luyện đồng thời nhiều bộ phận trên các mẫu bootstrap khác nhau rồi tổng hợp kết quả), Boosting (huấn luyện nối tiếp các bộ phận, mỗi bộ phận sau chú trọng vào các mẫu bị phán đoán sai trước đó), và Stacking (tổng hợp dự đoán của nhiều bộ phận thông qua một bộ học cấp trên). Trong địa hạt y khoa, kiến trúc tập thể tỏ ra đặc biệt giá trị vì mỗi bộ phận thành phần có thể nắm bắt các khía cạnh khác nhau của dữ liệu bệnh nhân; khi tổng hợp lại sẽ cho cái nhìn toàn diện và đáng tin hơn.

Đáng chú ý, Boosting (tiêu biểu như AdaBoost, Gradient Boosting hay XGBoost) thường được ưu tiên trong các bài toán lệch phân bố nhờ cơ chế tự động dồn sự chú ý vào các mẫu khó (thường chính là các mẫu yếu thế bị phán đoán sai) ở các vòng lặp sau. Một biến thể đáng chú ý là SMOTEBoost do Chawla và đồng nghiệp đề xuất năm 2003, tích hợp trực tiếp bước sinh mẫu SMOTE vào từng vòng lặp Boosting. Cụ thể, thay vì chỉ đơn thuần tăng trọng số cho các mẫu yếu thế bị phán đoán sai như AdaBoost truyền thống, SMOTEBoost sẽ phát sinh thêm mẫu yếu thế trước mỗi lượt huấn luyện bộ phận kế tiếp. Cơ chế này giúp thay đổi phân bố

trọng số một cách gián tiếp có lợi cho nhóm yếu thế, hướng mô hình tập trung mạnh hơn vào nhóm này. Kết quả thực nghiệm cho thấy SMOTEBoost đem lại bước cải thiện rõ rệt về năng lực dự báo nhóm ít gặp và nâng cao chỉ số F-measure trên nhiều tập dữ liệu có mức lệch cao. SMOTEBoost là minh chứng điển hình cho sức mạnh của việc kết hợp tăng mẫu với kiến trúc tập thể (cụ thể là Boosting) nhằm xử lý bài toán dữ liệu lệch phân bố.

Bên cạnh Boosting, chiến lược Bagging cũng sở hữu các biến thể thích ứng với dữ liệu lệch, điển hình như Balanced Random Forest – trong đó mỗi cây quyết định trong "khu rừng" được huấn luyện trên tập con dữ liệu đã được cân bằng (thường bằng cách giảm mẫu nhóm áp đảo hoặc tăng mẫu nhóm yếu thế trước khi huấn luyện từng cây). Cách làm này đảm bảo mỗi bộ phận thành phần "nhìn" dữ liệu một cách công bằng hơn; từ đó khi bỏ phiếu đa số ở giai đoạn tổng hợp cuối cùng, kết quả cũng bớt thiên lệch so với Random Forest tiêu chuẩn. Ngoài ra còn có EasyEnsemble và BalanceCascade – những thuật toán phối hợp nhiều vòng giảm mẫu với kiến trúc tập thể nhằm từng bước loại bỏ xu hướng thiên vị của nhóm áp đảo, đã được áp dụng thành công trong nhiều bài toán y sinh có dữ liệu lệch nghiêm trọng. Chẳng hạn, EasyEnsemble tạo nhiều tập con từ nhóm áp đảo rồi huấn luyện một bộ phân loại trên mỗi tập (kết hợp với toàn bộ nhóm yếu thế), sau đó tổng hợp các bộ phân loại; còn BalanceCascade thì loại bỏ dần các mẫu áp đảo được phân loại đúng qua các vòng bộ phân loại nối tiếp, nhằm dồn nguồn lực vào các mẫu áp đảo "khó" hơn. Tổng quan, kiến trúc tập thể cung cấp khung linh hoạt để lồng ghép các kỹ thuật xử lý lệch (như tăng/giảm mẫu) cùng với quá trình học mô hình, nhằm đạt hiệu năng tối ưu.

3. Chiến lược smote-ensemble cho nhận diện sớm biến chứng mắt

SMOTE-Ensemble không đơn thuần là một cụm từ chuyên môn mà đại diện cho một khung phương pháp luận toàn diện: phối hợp kỹ thuật tái cân bằng dữ liệu thông qua SMOTE với các kiến trúc học máy tập thể nhằm kiến tạo hệ thống nhận diện sớm biến chứng mắt do tiểu đường với hiệu suất vượt trội. Dựa trên nền tảng phân tích đã trình bày, quy trình triển khai chiến lược SMOTE-Ensemble có thể phác họa theo các bước sau:

Bước 1 - Tập hợp và chuẩn bị nguồn dữ liệu: Bao gồm các thông tin lâm sàng của người bệnh tiểu đường (tuổi, giới, khoảng thời gian mắc bệnh, chỉ số glucose huyết HbA1c, huyết áp, chức năng thận và các thông số sinh hóa khác), kèm theo nhãn phân loại xác định bệnh nhân đã hoặc chưa xuất hiện biến chứng mắt (dựa trên kết quả khám chuyên khoa mắt định kỳ hoặc kết luận của bác sĩ nhãn khoa). Đặc điểm nổi bật của nguồn dữ liệu này là mức độ lệch cao: số ca chưa biến chứng áp đảo số ca đã biến chứng.

Bước 2 - Tiền xử lý và trích xuất đặc trưng: Tiến hành làm sạch dữ liệu, xử lý các giá trị khuyết, và lựa chọn các đặc trưng có mối liên hệ mạnh nhất với nguy cơ biến chứng mắt (căn cứ vào y văn: ví dụ khoảng thời gian mắc bệnh kéo dài, HbA1c tăng cao, tăng huyết áp, bệnh thận do tiểu đường là những yếu tố nguy cơ chính). Bước này giúp mô hình hướng sự chú ý vào những tín hiệu quan trọng, giảm thiểu nhiễu, đồng thời thu gọn số chiều dữ liệu nhằm tạo điều kiện thuận lợi cho quá trình sinh mẫu SMOTE.

Bước 3 - Vận dụng SMOTE để tái cân bằng tập huấn luyện: Trên tập huấn luyện, triển khai thuật toán SMOTE nhằm phát sinh thêm các mẫu bệnh nhân đã biến chứng mắt (nhóm yếu thế) – các mẫu tổng hợp này được sinh ra dựa trên các bệnh nhân thực có trong nhóm. Tỷ lệ tăng mẫu có thể được điều chỉnh phù hợp với mức độ lệch của dữ liệu; ví dụ có thể nâng số mẫu nhóm yếu thế lên mức tương đương 50–100% nhóm áp đảo. Lưu ý quan trọng: chỉ áp dụng SMOTE trên tập huấn luyện (không áp dụng trên tập kiểm tra/đánh giá) nhằm tránh hiện tượng rò rỉ thông tin. Kết quả thu được là một tập dữ liệu huấn luyện đã được cân bằng hoặc gần cân bằng giữa hai nhóm.

Bước 4 - Huấn luyện kiến trúc tập thể: Lựa chọn kiến trúc tập thể phù hợp. Các thử nghiệm thực tế cho thấy Random Forest và XGBoost là hai lựa chọn hàng đầu trong bối cảnh dự báo y khoa nhờ sự kết hợp giữa năng lực dự báo mạnh mẽ và khả năng giải trình ở mức độ nhất định (Random Forest có thể cung cấp thông tin về tầm quan trọng của các đặc trưng) [8]. Với tập dữ liệu đã được tái cân bằng bằng SMOTE, tiến hành huấn luyện kiến trúc tập thể. Nhờ dữ liệu đã cân đối, mô hình có điều kiện học được cả hai nhóm một cách hiệu quả, giảm thiểu nguy cơ bỏ sót nhóm yếu thế.

Bước 5 - Đánh giá và tinh chỉnh: Đánh giá hiệu năng mô hình trên tập kiểm tra độc lập (hoặc sử dụng cross-validation) thông qua các chỉ số như AUC, Accuracy, và đặc biệt quan trọng là Recall (tỷ lệ phát hiện đúng) đối với nhóm biến chứng cùng Precision (độ chính xác dự đoán dương tính). Trong bối cảnh sàng lọc y tế, tỷ lệ phát hiện đúng cao được đặt ưu tiên nhằm hạn chế tối đa việc bỏ sót bệnh nhân thực sự mắc bệnh [8], đồng thời vẫn cần duy trì độ chính xác ở mức hợp lý để tránh tình trạng báo động giả quá nhiều. Nếu kết quả chưa đạt yêu cầu, có thể điều chỉnh các tham số của SMOTE (như số hàng xóm k, tỷ lệ tăng mẫu) hoặc thử nghiệm các chiến lược tập thể khác.

Thông qua việc phối hợp hai thành phần – SMOTE và kiến trúc tập thể – chiến lược SMOTE-Ensemble hướng đến việc khai thác đồng thời lợi thế của cả hai: (i) Cải thiện cấu trúc phân bố dữ liệu giúp mô hình nắm bắt được các tín hiệu đặc trưng của nhóm biến chứng (nhờ SMOTE), (ii) Nâng

cao độ chính xác tổng thể và tính ổn định của mô hình (nhờ sức mạnh tổng hợp của kiến trúc tập thể). Nhiều công trình nghiên cứu đã xác nhận hiệu quả của hướng tiếp cận này. Khan và cộng sự (2024) ghi nhận rằng các kiến trúc tập thể khi được tích hợp với kỹ thuật làm giàu dữ liệu (data augmentation) như SMOTE thường đạt kết quả vượt trội trong các bài toán nhóm ít gặp [9]. Trong lĩnh vực nhãn khoa, một số framework tiên tiến đã triển khai SMOTE-Ensemble. Chẳng hạn, SMOTE-Ensemble từng được đề xuất trong một giải pháp phân loại mức độ DR, phối hợp sinh mẫu ảnh võng mạc bằng GAN với tập hợp các mô hình học sâu, đạt mức cải thiện độ chính xác phân loại từ ~89,5% lên ~96,1% [10]. Đối với dữ liệu dạng bảng (thông tin lâm sàng bệnh nhân), Younseo Jang (2025) cũng phát triển kiến trúc tập thể dựa trên rừng ngẫu nhiên kết hợp tăng mẫu (SMOTE) và giảm mẫu, đạt AUC = 0,9227 trong dự báo sớm nguy cơ tiểu đường – vượt trội so với các mô hình đơn lẻ truyền thống. Mặc dù bài toán có sự khác biệt, kết quả này củng cố niềm tin rằng chiến lược tương tự hoàn toàn có thể áp dụng hiệu quả cho dự báo sớm biến chứng mắt trên dữ liệu bệnh nhân.

Tóm lại, chiến lược SMOTE-Ensemble tận dụng được cả năng lực mở rộng dữ liệu lẫn tổng hợp tri thức từ nhiều mô hình. Đây là một hướng tiếp cận đầy triển vọng để xử lý bài toán nhận diện sớm biến chứng mắt do tiểu đường, vốn đòi hỏi mô hình vừa có tỷ lệ phát hiện đúng cao (nhận ra được các ca bệnh ít gặp) vừa đảm bảo độ chính xác (hạn chế báo động giả không cần thiết).

4. Kết quả các nghiên cứu

Để đánh giá hiệu quả của hướng tiếp cận SMOTEEnsemble, chúng tôi tổng hợp kết quả từ nhiều nghiên cứu gần đây. Bảng 1 liệt kê một số công trình tiêu biểu cùng dữ liệu và mô hình sử dụng, kèm theo các chỉ số chất lượng quan trọng là AUC, F1-score và Recall (độ nhạy) của mô hình. Các nghiên cứu bao gồm cả mô hình dự đoán DR trên dữ liệu lâm sàng và các mô hình dự báo biến chứng đái tháo đường nói chung, nhằm so sánh hiệu năng giữa các cách tiếp cận khác nhau.

Qua bảng trên có thể thấy, các mô hình kết hợp kỹ thuật cân bằng dữ liệu và ensemble nói chung đạt kết quả rất tích cực. Chẳng hạn, mô hình XGBoost của X. Wan *et al.* đạt AUC xấp xỉ 0,97 – rất cao so với mặt bằng chung [2] Tương tự, nghiên cứu của Jang *et al.* với SMOTE+Ensemble cũng đạt AUC > 0,92, cải thiện đáng kể so với các mô hình đơn lẻ trước đó. Ở phía ngược lại, nếu không xử lý mất cân bằng, mô hình có thể chỉ đạt AUC khoảng 0,75–0,80 (như kết quả của Ogunyemi *et al.* hoặc Yang *et al.*). Đặc biệt, việc nâng recall (độ nhạy) cho lớp biến chứng là ưu tiên hàng đầu – nhiều nghiên cứu cho thấy SMOTE giúp tăng recall lên mức ~90% hoặc cao hơn, đồng thời giữ F1-score và AUC ở mức chấp nhận được [2], [8].

Cũng cần lưu ý rằng các mô hình học sâu trên ảnh võng mạc (như của Gulshan *et al.*) có thể đạt độ chính xác rất cao (AUC ~0,99)[10]. Tuy nhiên, những mô hình này đòi hỏi dữ liệu ảnh số lượng lớn và hạ tầng tính toán mạnh, không phải lúc nào cũng sẵn có. Trong khi đó, các mô hình SMOTEEnsemble dựa trên dữ liệu lâm sàng có ưu thế là tận dụng được dữ liệu sẵn có rộng rãi (xét nghiệm máu, hồ sơ bệnh án) và triển khai dễ dàng trong hệ thống quản lý bệnh nhân. Do vậy, hai hướng tiếp cận có thể bổ trợ cho nhau: mô hình trên dữ liệu bảng giúp sàng lọc ban đầu để chọn ra nhóm nguy cơ cao, sau đó mô hình học sâu trên ảnh sẽ phân tích chi tiết tổn thương đáy mắt cho nhóm này.

5. Thảo luận

Các kết quả phân tích ở trên cho thấy việc phối hợp Những kết quả phân tích trình bày ở trên cho thấy việc phối hợp kỹ thuật tăng mẫu (tiêu biểu là SMOTE) với các kiến trúc học máy tập thể đã đem lại những bước tiến đáng kể trong bài toán nhận diện sớm biến chứng mắt do tiểu đường cũng như các biến chứng khác của rối loạn chuyển hóa glucose. Khung phương pháp SMOTE-Ensemble nhắm tới việc khắc phục hai rào cản cốt lõi vốn gây khó khăn cho các thuật toán học máy truyền thống: thứ nhất là tình trạng lệch phân bố giữa các nhóm ở mức nghiêm trọng, và thứ hai là mối quan hệ phi tuyến, đa chiều giữa các yếu tố nguy cơ với biến chứng.

Nhìn từ góc độ dữ liệu, chiến lược tăng mẫu thông qua SMOTE đã chứng tỏ năng lực nâng cao tỷ lệ phát hiện đúng đối với các trường hợp biến chứng vốn chiếm tỷ trọng thấp. Nhờ cơ chế phát sinh thêm các mẫu tổng hợp dựa trên thông tin từ bệnh nhân thực, thuật toán học máy không còn xu

Bảng 1. Kết quả dự đoán biến chứng DR/đái tháo đường trong các nghiên cứu tiêu biểu (AUC: diện tích dưới đường cong ROC; F1: điểm F1; Recall: độ nhạy phát hiện biến chứng).

Nghiên cứu (năm)	AUC	F1-score (%)	Recall (%)
Ogunyemi et al. (2021) – Mô hình DNN trên EHR (xử lý mất cân bằng bằng undersampling) [4]	0,80	–	72,2
Wan *et al.* (2025) – Mô hình XGBoost dự đoán DR dựa trên xét nghiệm thường quy [2]	0,831	75,2	75,4
Gulshan *et al.* (2016) – Mạng CNN phát hiện DR trên ảnh võng mạc (Google Research) [10]	0,991	–	90 (Se) / 98 (Sp)

(*Chú thích: Dấu "–" nghĩa là không được báo cáo cụ thể trong nghiên cứu. Se/Sp: Sensitivity/Specificity.*)

hướng "phớt lờ" nhóm yếu thế, qua đó hạn chế đáng kể tình trạng bỏ sót ca bệnh. Trong bối cảnh sàng lọc y tế, một mô hình đạt recall cao đồng nghĩa với việc hầu như toàn bộ bệnh nhân có tổn thương mắt đều được ghi nhận để theo dõi tiếp – đây chính là mục tiêu tối quan trọng xét về khía cạnh lâm sàng. Mặc dù đánh đổi có thể là sự sụt giảm một phần precision (xuất hiện nhiều ca dương tính giả hơn), điều này hoàn toàn có thể chấp nhận được trong quy trình sàng lọc, với điều kiện các bước xác nhận sau đó (chẳng hạn bác sĩ chuyên khoa khám xác nhận) được thực hiện để loại trừ các cảnh báo sai.

Nhìn từ góc độ mô hình, kiến trúc học tập thể (ensemble learning) đóng vai trò như một lớp phòng vệ chống lại sự bất định và nguy cơ khớp quá của các mô hình đơn lẻ. Các bộ phận thành phần trong kiến trúc tập thể – có thể là tập hợp nhiều cây quyết định trong Random Forest, hoặc chuỗi các booster nối tiếp trong XGBoost – sẽ hỗ trợ bù đắp cho những điểm yếu của nhau. Đặc tính này đặc biệt có giá trị khi xử lý dữ liệu phức tạp và nhiều nhiễu, như hồ sơ y tế bệnh nhân với nhiều đặc trưng có mối tương quan lẫn nhau. Kiến trúc tập thể còn cho phép tích hợp các thuật toán thuộc nhiều họ khác nhau (heterogeneous ensemble), ví dụ kết hợp cả mô hình tuyến tính lẫn phi tuyến, nhằm tận dụng tối đa ưu điểm của từng loại. Kết quả là mô hình tổng hợp cuối cùng vừa đạt được độ chính xác tổng thể cao hơn, vừa thể hiện tính ổn định tốt hơn khi triển khai trên dữ liệu mới [8].

Một khía cạnh đáng lưu ý là mô hình SMOTE-Ensemble vận hành trên dữ liệu lâm sàng bệnh nhân có thể đóng vai trò bổ trợ cho các hệ thống AI phân tích hình ảnh võng mạc. Ở thời điểm hiện tại, các thuật toán học sâu xử lý ảnh có khả năng đạt độ chính xác rất cao trong phát hiện DR trên hình ảnh đáy mắt (như thuật toán của Google đạt AUC ~0,99) [10]. Tuy vậy, để triển khai ở quy mô rộng, giải pháp dựa trên ảnh đòi hỏi đầu tư hệ thống camera chụp đáy mắt cùng đội ngũ bác sĩ nhãn khoa đọc và xác nhận kết quả. Trong khi đó, một mô hình SMOTE-Ensemble vận hành trên dữ liệu hành chính và kết quả xét nghiệm có sẵn từ hồ sơ bệnh nhân có thể đảm nhiệm chức năng như một công cụ sàng lọc tự động ở giai đoạn đầu. Hệ thống sẽ liên tục rà soát hồ sơ để nhận diện những bệnh nhân có nguy cơ cao (căn cứ vào các yếu tố như khoảng thời gian mắc bệnh kéo dài, kiểm soát glucose kém, đã có biến chứng thận, v.v.) [8]. Sau đó, chỉ những bệnh nhân thuộc nhóm này mới được ưu tiên mời đến chụp ảnh võng mạc sớm. Cách triển khai này giúp tối ưu hóa nguồn lực y tế, dồn nỗ lực khám chuyên sâu cho nhóm thực sự cần thiết, đồng thời giảm thiểu nguy cơ bỏ sót bệnh nhân.

Mặc dù mang nhiều triển vọng, chiến lược SMOTE-Ensemble cũng tồn tại một số điểm hạn chế cần lưu ý. Thứ nhất, việc phát sinh quá nhiều mẫu tổng hợp có thể dẫn đến hiện tượng khớp quá nếu kiến trúc mô hình quá phức tạp so với bản chất dữ liệu thực. Bởi vậy, cần cân nhắc kỹ lưỡng khi lựa chọn tỷ lệ tăng mẫu và nên phối hợp với các kỹ thuật như cross-validation hoặc early stopping nhằm phòng ngừa overfitting. Thứ hai, SMOTE vận hành dựa trên giả định rằng các hàng xóm gần trong không gian đặc trưng mang ý nghĩa tương đồng – giả định này có thể không chính xác nếu dữ liệu chứa nhiều nhiễu hoặc các đặc trưng được lựa chọn chưa phù hợp. Giải pháp khắc phục là nên thực hiện bước chọn lọc đặc trưng (feature selection) hoặc giảm chiều (PCA) trước khi áp dụng SMOTE nhằm loại bỏ nhiễu [8]. Thứ ba, về phương diện mô hình tập thể, mặc dù Random Forest và XGBoost thể hiện hiệu quả cao, nhưng khả năng giải trình mô hình (model interpretability) vẫn là một thách thức đáng kể. Trong môi trường y khoa, đội ngũ bác sĩ thường mong muốn hiểu rõ yếu tố nào đóng vai trò quyết định. Các công cụ như SHAP value hoặc LIME có thể được tích hợp để giải trình mức độ đóng góp của từng đặc trưng trong quá trình ra quyết định của mô hình [7]. Việc bổ sung khả năng giải trình sẽ góp phần gia tăng sự tin tưởng và chấp nhận của cộng đồng y tế đối với các mô hình AI.

6. Kết luận

Nhận diện sớm biến chứng mắt ở cộng đồng người bệnh tiểu đường là một nhiệm vụ mang ý nghĩa thiết yếu, góp phần giảm thiểu nguy cơ mất đi vĩnh viễn khả năng nhìn đồng thời giảm bớt gánh nặng cho hệ thống y tế. Rào cản lớn nhất của nhiệm vụ này bắt nguồn từ đặc điểm dữ liệu y khoa thường có phân bố lệch nghiêm trọng, trong đó số ca biến chứng chỉ chiếm một tỷ trọng rất nhỏ so với tổng số bệnh nhân. Thông qua quá trình khảo sát và phân tích, có thể khẳng định rằng việc phối hợp SMOTE để tái cân bằng dữ liệu cùng với kiến trúc học tập thể (ensemble learning) để gia tăng sức mạnh dự báo là một giải pháp hiệu quả cho vấn đề nêu trên. Kiến trúc SMOTE-Ensemble hỗ trợ thuật toán học máy nhận dạng tốt hơn các dấu hiệu tiền triệu của biến chứng mắt, cải thiện đáng kể tỷ lệ phát hiện đúng trong việc nhận diện bệnh nhân nguy cơ cao mà không gây suy giảm đáng kể độ chính xác tổng thể. Các công trình nghiên cứu thực nghiệm đã ghi nhận mô hình ứng dụng SMOTE-Ensemble đạt kết quả vượt trội về AUC, F1-score, và đặc biệt là nâng cao recall cho nhóm biến chứng so với các mô hình truyền thống [7], [8].

Chiến lược SMOTE-Ensemble thể hiện tính khả thi cao để triển khai trong thực tiễn lâm sàng: nguồn dữ liệu đầu vào là các thông tin y tế sẵn có và thuật toán học máy hoàn toàn có thể được tích hợp vào hệ thống quản lý bệnh nhân hiện hành. Khi đưa vào vận hành, mô hình sẽ liên tục phân tích dữ liệu mới nhập và đưa ra cảnh báo sớm cho đội ngũ y bác sĩ về những bệnh nhân tiểu đường có khả năng đã hoặc sắp xuất hiện biến chứng mắt, từ đó hỗ trợ quyết định chỉ định khám chuyên khoa mắt kịp thời. Điều này đặc biệt có giá trị tại những vùng miền khan hiếm nhân lực nhãn khoa – công nghệ

AI sẽ đóng vai trò như một "trợ lý thông minh" hỗ trợ sàng lọc ở giai đoạn đầu. Trong tương lai, SMOTE-Ensemble có thể được mở rộng tích hợp với các kỹ thuật tiên tiến khác như học chuyển giao (transfer learning) từ các mô hình phân tích ảnh võng mạc, hoặc học sâu đa nguồn (multimodal learning) kết hợp đồng thời cả ảnh võng mạc và dữ liệu lâm sàng để gia tăng thêm độ chính xác. Tuy vậy, dù triển khai theo hướng nào, nguyên tắc cốt lõi vẫn là đảm bảo mô hình không bị thiên vị và có khả năng nhận diện được những "tín hiệu nhỏ giữa biển dữ liệu" – chính là các bệnh nhân đang âm thầm mang trong mình biến chứng. Với hướng tiếp cận SMOTE-Ensemble, chúng ta đang tiến thêm một bước trong hành trình ứng dụng hiệu quả trí tuệ máy vào công tác chăm sóc người bệnh tiểu đường, phát hiện sớm biến chứng mắt và góp phần gìn giữ khả năng nhìn cho hàng triệu người.

Tài liệu tham khảo

[1] S. Yadav, "A Comparative Analysis of Sampling Techniques for Imbalanced Datasets in Machine Learning," *International Journal of Innovative Research and Development*, vol. 7, Sep. 2021, doi: 10.5281/zenodo.14203644.

[2] X. Wan *et al.*, "Predicting diabetic retinopathy based on routine laboratory tests by machine learning algorithms," *Eur J Med Res*, vol. 30, no. 1, p. 183, 2025, doi: 10.1186/s40001-025-02442-5.

[3] D. Yan, X. Li, Y. Wang, and Z. Cai, "Optimized prediction of diabetes complications using ensemble learning with Bayesian optimization: a cost-efficient laboratory-based approach," *Front Endocrinol (Lausanne)*, vol. 16-2025, 2025, doi: 10.3389/fendo.2025.1593068.

[4] O. I. Ogunyemi *et al.*, "Detecting diabetic retinopathy through machine learning on electronic health record data from an urban, safety net healthcare system," *JAMIA Open*, vol. 4, no. 3, p. ooab066, Aug. 2021, doi: 10.1093/jamiaopen/ooab066.

[5] Georgedouzas, "georgedouzas / imbalanced-learn-extra Public," https://github.com/georgedouzas/imbalanced-learn-extra.

[6] S. Yadav, "A Comparative Analysis of Sampling Techniques for Imbalanced Datasets in Machine Learning," *International Journal of Innovative Research and Development*, vol. 7, Jul. 2021, doi: 10.5281/zenodo.14203644.

[7] X. Wan *et al.*, "Predicting diabetic retinopathy based on routine laboratory tests by machine learning algorithms," *Eur J Med Res*, vol. 30, no. 1, p. 183, 2025, doi: 10.1186/s40001-025-02442-5.

[8] A. Khan, O. Chaudhari, and R. Chandra, "A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation," *Expert Syst Appl*, vol. 244, p. 122778, Jul. 2024, doi: 10.1016/j.eswa.2023.122778.

[9] S. Naik, D. Kamidi, S. Govathoti, R. Cheruku, and A. Reddy, "RETRACTED ARTICLE: Efficient diabetic retinopathy detection using convolutional neural network and data augmentation," *Soft comput*, vol. 28, p. 617, Jun. 2023, doi: 10.1007/s00500-023-08537-7.

[10] V. Gulshan *et al.*, "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs," *JAMA*, vol. 316, no. 22, pp. 2402–2410, Dec. 2016, doi: 10.1001/jama.2016.17216.

SMOTE-ENSEMBLE: A REVIEW OF DATA-BALANCING TECHNIQUES AND HYBRID MACHINE LEARNING MODELS FOR EARLY PREDICTION OF DIABETIC RETINOPATHY COMPLICATIONS

Abstract:

Diabetic retinopathy (DR) is one of the most common microvascular complications of diabetes and the leading cause of vision loss worldwide, yet its early stages often produce no noticeable symptoms. Consequently, largescale fundus image screening programs face challenges due to the inherent class imbalance in DR datasets: severe cases requiring urgent intervention are scarce compared to mild or nonDR images. To address this, we provide a comprehensive overview of SMOTE ensemble strategies to enhance detection sensitivity for underrepresented classes. First, we analyze DR's pathophysiological progression and dataset characteristics, demonstrating how imbalance reduces model recall. We then detail SMOTE and its variants—including BorderlineSMOTE, ADASYN, and Geometric SMOTE—highlighting controlled synthetic minority oversampling [1]. Next, we review ensemble learning frameworks (Bagging, Boosting, Voting, Stacking) and their integration with SMOTE, with emphasis on the SMOTEBoost algorithm and recent refinements. Synthesizing results from over thirty studies, we show that SMOTE ensemble methods yield 5–18 % improvements in AUC, F1 score, and recall for severe DR detection [2].

Finally, we discuss current limitations and propose future research directions—such as GANbased augmentation, multitask learning, and interpretable model design—to accelerate clinical deployment.

Keywords: *SMOTE Ensemble ,Ensemble Learning, Borderline-SMOTE, ADASYN, Geometric, imbalance data.*